Evaluating Large Language Models in Psychological Research: A Guide for Reviewers

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

[1]Department of Psychology, University of Southern California

[2]Brain and Creativity Institute, University of Southern California

[3]Department of Computer Science, University of Southern California

[4]Department of Marketing, University of Illinois Chicago

[5]Department of Marketing, University of Chicago

Author Note

Suhaib Abdurahman https://orcid.org/0000-0001-5615-0129

Correspondence concerning this article should be addressed to Suhaib Abdurahman, Department of Psychology, University of Southern California, 501 Seeley G. Mudd Building, 3620 McClintock Ave, Los Angeles, CA 90089. E-mail: sabdurah@usc.edu

Abstract

Large Language Models (LLMs) are being increasingly used in scientific research, be it to analyze data, generate synthetic data, or even to write scientific papers. This trend necessitates that journal reviewers are able to evaluate the quality of works that utilize LLMs. We provide reviewers of psychological research with a comprehensive guide on evaluating research that uses LLMs, examining their dual roles of automating data processing and simulating human data. Essential considerations for reviewers are highlighted, focusing on the evaluation of methodological rigor, the importance of replicability, and the validity of results when employing LLMs. 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 non-deterministic and continuously evolving nature of these models. By providing a framework for critical review, this guide aims to ensure high-quality, innovative research within the evolving landscape of psychological studies utilizing LLMs.

Keywords: psychology, large language models, natural language processing, psychological text analysis, synthetic data

Evaluating Large Language Models in Psychological Research: A Guide for Reviewers

Large Language Models (LLMs) are becoming increasingly prevalent in scientific research. Liang et al. (2024) found that between 6.3 and 17.5% of scientific publications since 2020 use LLMs in their writing, with a steady increase in utilization rates. LLMs also play an increasingly important role in psychological research specifically. A 2024 study by Ke et al. (2024), presenting a comprehensive overview of over 100 recent works, demonstrates the increasing adoption of LLMs across diverse sub-fields of psychology such as cognitive, social, cultural, clinical, and developmental psychology. Currently, LLMs in psychological research are primarily used in two capacities: LLMs can automate costly tasks, such as coding qualitative and free-response data (Amin et al., 2023; Rathje et al., 2023; Törnberg, 2023; Ziems et al., 2024), and some researchers have even suggested that LLMs can simulate complex social and cognitive phenomena (Aher et al., 2023; Argyle et al., 2023; Bai et al., 2023; Coda-Forno et al., 2023; Dillion et al., 2023; He et al., 2024; Horton, 2023; Park et al., 2023; Suri et al., 2024). These developments necessitate that academic journal reviewers understand how to appropriately use LLMs to ensure high-quality and replicable research. In this article, we highlight important factors that reviewers should consider as they review papers that use LLMs. We provide practical advice on how to evaluate these works, focusing mainly on ensuring that robust inferences can be made from the experiments and simulations. In this context, we cover currently appropriate ways of using LLMs in psychological research, with a particular focus on the replication and validity of results. See Table 1, for an introduction to key terms used in relation to LLMs.

Recently, LLMs have been used to simulate human-like behaviors and social interactions (e.g., Park et al., 2023) or to simulate human responses to psychological questionnaires and vignettes (e.g., Dillion et al., 2023). Although these applications are in their infancy, they represent promising developments that could facilitate innovative

Table 1

Overview of technical terms.

Term	**Explanation**
Generative AI	A type of AI that focuses on creating content, such as text, images, or music, usually trained to resemble human-generated content.
Autoregressive Models	Models that predict future values based on past values, commonly used in time series forecasting and language modeling.
Large Language Model (LLM)	An autoregressive model for generating and understanding human language. LLMs can automate complex tasks, such as interpreting text data.
Embedding	A vector representation of, e.g., text data in a continuous, real number vector space.
Fine-Tuning	The process of taking a pre-trained model and further training it on a specialized dataset to improve its accuracy on a given task.
Prompt	The input to an LLM. It usually contains instructions to generate a specific output or perform a task. The prompt design is crucial, as it influences the LLM's autoregressive content generation.
Zero-Shot, Few-Shot, Many-Shot	Learning paradigms where a model performs tasks with no prior examples (zero-shot), few examples (few-shot), or many examples (many-shot).
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.
Application Programming Interface (API)	An API is a set of definitions and protocols that allows software applications to communicate with each other. In the context of LLMs, they allow a user's program to access the model.

research methods particularly for preliminary stages of psychological research (e.g., piloting of studies, and assisting with power analyses). However, for now, most uses of LLMs in psychological research involve evaluating human responses and observational data rather than producing primary data. Free response data is a valuable resource for understanding human thoughts and behaviors (Ericsson & Moxley, 2019), but coding this kind of data to make it suitable for analysis can be labor-intensive, especially for large samples of observational data (i.e., user reviews from websites, social media posts, news articles). LLMs offer a simpler and less labor-intensive way of coding this type of data (Chiang & Lee, 2023; Gilardi et al., 2023; Naismith et al., 2023; Rathje et al., 2023; Tabone & de Winter, 2023). LLMs are gaining in popularity, compared to other methods of natural language processing (NLP) because they can allow researchers to automatically process data more conveniently. Researchers may use the same LLM for various tasks, without task-specific data collection and fine-tuning. Often, the LLM's performance is on par with other NLP methods and in some cases even significantly better (Abdurahman et al., 2023; Rathje et al., 2023).

While LLMs are gaining in popularity, they should not be the automatic choice for all language processing tasks. In some cases, LLMs may require significant customization through fine-tuning or more complex prompting strategies that include examples and explanations to achieve competitive performance (Abdurahman et al., 2023; Brown et al., 2020). Additionally, established techniques such as dictionary-based approaches and approaches based on smaller language models' embeddings can be easier to interpret, replicate, and, in some cases, even outperform LLMs. For example, there are several standardized dictionary-based techniques for general linguistic evaluation (e.g., Pennebaker et al., 2007) sentiment analysis tools (e.g., VADER: Hutto & Gilbert, 2014), and assessing moral content (e.g., MFD: Graham et al., 2009). Dictionary approaches are theory-based and interpretable but usually outperformed by language model based approaches. Smaller language models, such as BERT (Devlin et al., 2018) can be hosted on consumer grade hardware, and can be fine-tuned such that they often outperform LLMs, especially LLMs in zero-shot settings (Abdurahman et al., 2023). They also offer greater replicability and transparency because researchers have access to the model weights and the model's underlying representations of text (i.e., embeddings). This allows researchers to interpret language model outputs, such as through post-hoc explanations (e.g., examining the impact of phrases in the input on the model output) or through directly examining the model layers' activation (e.g., Kennedy et al., 2020, 2021; Serrano & Smith, 2019). Lastly, hybrid models that combine language models' accuracy with traditional methods' interpretability and replicability—for example, Distributed Dictionary Representations (DDR; Garten et al., 2018) and Contextual Context Representations (CCR; Atari et al., 2023)—are also promising. DDR integrates dictionary content with language models for text analysis, while CCR integrates psychological instruments (e.g., questionnaire items) with language models, showing promising results in tasks such as extracting values and beliefs from free response data (Abdurahman et al., 2023; Atari et al., 2023). However, many benefits of smaller language models (e.g., BERT's high performance after fine-tuning, or easy access to its embeddings) can be achieved with LLMs but they might be significantly more complicated, require technical know-how (e.g., for custom implementations), and access to the model that only few (e.g., open-source models) provide. Researchers should keep these considerations in mind when deciding whether to use an LLM or its alternatives.

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. In the following, we present a set of guidelines to help reviewers identify potential issues with replication and validation of research that uses LLMs and potential ways of addressing them. Importantly, reviewers should look at the following guidelines from a holistic perspective. The main goal should be to ensure that authors' methodology is sound and that their conclusions are justified. Reviewers, with their domain expertise, play a crucial role in ensuring these guidelines are applied appropriately based on the specific context of each study. Simply put, the more central the LLM use is to the research question, the more strongly should reviewers consider the recommendations. For example, if an LLM is only used for a minor task like automating data coding, ensuring accuracy and replicability might be sufficient. However, when LLMs are used to draw conclusions about human behavior (e.g., by simulating human data) or their capabilities are directly studied, thorough validation, robustness checks, and replicability become crucial. Reviewers' responsibilities and workload should not differ significantly compared to non-LLM works. They mainly need to verify that all relevant information regarding the authors methods is provided (similar to how experimental designs are explained), that their methodological choices are justified and discussed (similar to how experimental choices have to be justified), and that the model outputs are validated and replicable (similar to how experimental works may require manipulation checks and validation with external data). See Figure 1 for a high-level overview of the key aspects of the reviewing process and see the appendix for a detailed checklist.

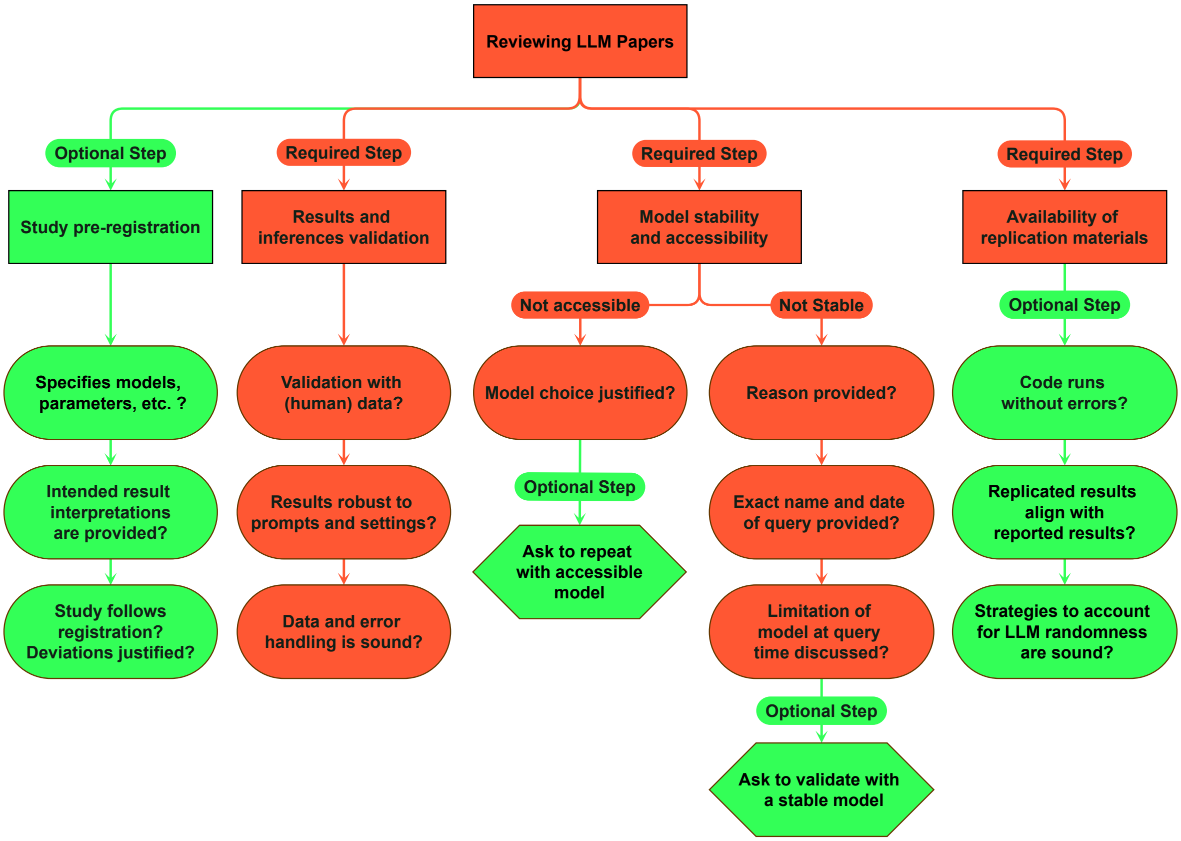
Ensuring Replicable Research with LLMs

Transparency and Accessibility of Materials

Most casual users of LLMs are accustomed to interacting with them through a chat interface like the one provided by OpenAI's ChatGPT. Where LLMs are used to process large amounts of data or where model settings must be controlled, a researcher will typically interact with the LLM through a program. These programs feed data, through an API, 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 the norm for other research that authors provide data, computer code, experimental instructions, materials, and procedures needed to replicate findings, reviewers of articles using LLMs should be provided materials necessary for them and others to replicate the findings in the article under review. This

Figure 1

Overview of the key reviewing considerations.


Reviewing LLM Papers
Optional Step
Required Step
Required Step
Required Step
Study pre-registration
Results and inferences validation
Model stability and accessibility
Availability of replication materials
Not accessible
Not Stable
Optional Step
Specifies models, parameters, etc. ?
Validation with (human) data?
Model choice justified?
Reason provided?
Code runs without errors?
Intended result interpretations are provided?
Results robust to prompts and settings?
Optional Step
Exact name and date of query provided?
Replicated results align with reported results?
Ask to repeat with accessible model
Study follows registration? Deviations justified?
Data and error handling is sound?
Limitation of model at query time discussed?
Strategies to account for LLM randomness are sound?
Optional Step
Ask to validate with a stable model

Note. First row of rectangles denotes the key considerations when reviewing LLMs. Red indicates required steps, green indicates optional steps. Subsequent rows of pills indicate the different steps for the respective topics. Capped pills indicate requests to authors.

includes the code and materials to run the studies and analyses. Importantly, reviewers need to be able to see the exact information that is fed to the LLM as well as any model settings. Ideally, reviewers would attempt to replicate the results using the materials and let failure to replicate the results or deviations inform their review decisions.

In summary, reviewers should make sure that authors are transparent with their methods, provide all relevant information and materials, and guarantee easy replication of their studies.

Dealing with Non-Determinism

LLMs pose unique challenges due to their non-deterministic nature. An LLM may code a piece of text differently each time it is asked to, even when using the same prompts and model setting. Reviewers should ensure that the reported findings are reliable and not coincidental. For example, authors could run the experiment repeatedly and report and discuss the variation in outcomes. 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 should be reported (e.g., means and standard deviations of numeric results, percentage of simulations that show a target behavior or other non-numeric observations). Importantly, reviewers who use an author's code and/or prompts to replicate results should anticipate minor variation between the reported and replicated results (akin to what would be expected in bootstrapping and other simulation-based approaches). However, significant discrepancies or replication failures should reduce confidence in the author's claims. In summary, reviewers should verify that strategies to ensure the reliability of LLM results are applied. They should pay special attention to how authors deal with randomness in LLM outputs and how they make sure that their results are not coincidental. See an overview of guidelines to ensure replicable LLM research in Table 2.

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. Specifically, the performance of the same model can even decrease over time due to updates (L. Chen et al., 2023) and newer models might underperform in some areas compared to previous versions (Achiam et al., 2023; Coyne & Sakaguchi, 2023; Rathje

Table 2

Reviewer guidelines to ensure replicable LLM research.

Potential Issue	**Strategy**	**Details**
Inaccessibility/ Intransparency	Full access to materials and documentation	Ensure authors provide all data, code, experimental instructions, and model settings to replicate the study findings.
Non-Determinism	Multiple codings and simulations	For various tasks, LLMs should be prompted repeatedly. Reports should include, e.g., means and standard deviations, or majority votes to showcase consistency or variability.
Model Updates	Document model version, use stable models, or use local open-source models	Check that authors have specified LLM version and query date. Recommend stable versions mitigating risks associated with model changes. Encourage local open-source models that can be shared with others.

et al., 2023). Owners of these models are not required to and typically do not disclose every and all changes or provide long-term access to previous versions. Results obtained from OpenAI's GPT-4 when an article is submitted may not replicate with GPT-4 months later when it goes to press, making research using these models difficult to build upon.

To ensure replicability, reviewers should, if possible, recommend that authors use a model that is stable over time and permanently available. Open-source models are increasingly available—like Meta's LLaMa family of models (Touvron et al., 2023), the open-science driven BLOOM (Le Scao et al., 2023), and some recent Mistral models (e.g., Mistral: Jiang et al., 2023)—and offer these advantages. These models underlie an ecosystem of open-source LLMs available through platforms such as HuggingFace and include various fine-tuned and optimized derivatives. They vary in their power, specialize in different tasks, and have different safety controls. Some of these models compete with state-of-the-art proprietary models such as GPT-3.5 or GPT-4. Note that the performance landscape for LLMs is continuously evolving, as new models are developed and existing ones are refined. Leaderboards such as HuggingFace's Chatbot Arena (https://chat.lmsys.org/?leaderboard) and Open LLM Leaderboard (https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard) provide current benchmarks for both open and closed-source LLMs. As of now, proprietary models have an edge in user adoption due to their convenience and ease of use through various commercial services (e.g., chat-interfaces, API) that allow users to utilize these models without much coding knowledge or familiarity with self-hosting language models. However, open-source models give researchers more control over the model they are using and allow for reproducibility in a way commercial models do not (e.g., because the model weights can be stored and will not change over time, the weights can be accessed and analyzed) and should thus be preferred to ensure replicability. Additionally, some companies such as Mistral offer APIs for open-source models (often at lower costs than proprietary models), increasing accessibility for researchers that cannot self-host. If authors use open-source models, reviewers should require that authors share the model weights (or link to platforms such as HuggingFace that store these weights) in addition to their codes and all model settings.

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 their choice of model that cannot guarantee replication, and reviewers should weigh this justification in their evaluation of the paper. Even when using current proprietary models, it is advisable to choose models 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 wasn't used, authors should ideally replicate with a current version of the model or a different (stable) model. Reviewers should furthermore recommend authors to disclose all details about the utilized model versions. This includes both the exact name of the model as well as the time-point of the data collection (e.g., "gpt-4-0125-preview" or "gpt-4-1106-preview", instead of only "gpt-4-turbo").

In summary, reviewers should require a statement about the replicability of the results that involves justification of the choice of model, including potential long-term changes to the models.

Validating Research that utilizes LLMs

In all cases, reviewers should require that authors explain how they assessed the reliability and validity of an LLM's responses. Researchers have many degrees of freedom when generating LLM outputs. Reviewers should thus require that authors 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. See an overview of the guidelines and recommended strategies in Table 3.

Confirming LLM Outputs with Human Data

Reviewers evaluating research using LLMs for classification should recommend that authors validate LLM responses to assess accuracy and bias. This can be done in either of two ways. First, the authors could refer to past literature that validated a model's performance on a given task. It should be noted 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 where a model is both stable over time and has been thoroughly validated so that its performance can simply be assumed without any test by the authors. However, this might change as LLMs improve, as more research and application shifts to open-source models (or as proprietary models provide permanently stable versions), and as more validation studies are published. Second, where direct past validation is lacking, authors should validate the model's performance for their respective studies. Like reporting interrater reliability across multiple human raters, coding tasks using LLMs should report model accuracy by comparing responses from the LLM and human coders (or any other ground-truth data) for a subsample of responses. For example, an author might have an LLM and three raters code free response data from 100 out of 1000 participant responses. If sufficiently high accuracy is reached, the author may then code the remaining 900 responses using only the LLM. Considering that authors have substantial flexibility in selecting the subsample to evaluate the model's performance, authors should ideally make these comparisons on a pre-defined and justified subsample of the data to guard against researchers oversampling responses where human and LLM coding match (analogous to "p-hacking" where researchers run multiple analyses and only report significant ones). Ideally, authors should also investigate whether the false classifications by the model correlate with relevant features of the text or task regarding the research question (e.g., ruling out that tweets by women, compared to men, are more frequently misclassified in a study that investigates gender differences). This is an important validation step for any predictive method, be it a generative LLM or a classical machine learning system such as KNN (Fix & Hodges, 1951).

In summary, reviewers should ensure that authors validate their model's accuracy and, ideally, ensure its biases do not systematically skew their results and subsequent inferences. This can be done by comparing its outputs with human or other ground-truth data, referring to recent literature on the same tasks, and thoroughly examining model choice and application.

Robustness of Prompts

One of the most fundamental choices in using an LLM is how an author instructs it to classify data (i.e., prompt the model). LLMs, like humans, can alter their responses based on prompt wording (Abdurahman et al., 2023; Fujita et al., 2022; Lu et al., 2021). Additionally, 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 based on the number of examples, and is often employed by researchers to increase model performance (Brown et al., 2020; Wang et al., 2020). However, this introduces yet another degree of freedom because the choice of examples, and even their order, can change the model's outputs (Lu et al., 2021). Therefore, authors should provide the specific prompts and examples that they used. Ideally, they would also justify their prompts (e.g., based on theory-driven considerations), or use strategies that test the robustness of prompting strategies, analogous to how psychologists account for differences in participant responses using stimulus sampling, randomization of question order, different scales, etc. 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).

In summary, reviewers should ensure that authors 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 towards desired, yet only coincidental, outcomes.

Model Settings

Parameters

Authors can tweak several settings when interacting with an LLM. For example, an important model parameter is the so-called "temperature". LLMs are autoregressive models that predict the probability of the next token (i.e., word) given the previous ones. The model then chooses the next token based on the probability distribution. A low temperature setting leads the model to choose the most likely token, while a high temperature leads the model to sample according to the probability distribution. Importantly, even though temperature is sometimes compared to creativity or even used to induce variance in model responses (Almeida et al., 2023; Atari et al., 2023; Davis et al., 2024; Zhao et al., 2024), it is not the same as the variance measured across participants that psychologists usually care about. Using a high temperature is affected by how sure the model is about a response (i.e., how “sharp” is the probability distribution over the possible outputs). It is thus more similar to the variance measured within a participant, i.e. asking the same person the same question multiple times (Abdurahman et al., 2023; Park et al., 2023). 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 tokens within a response or limit the length of model output. Reviewers should make sure that authors report and justify relevant model settings.

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 within a batch, or simply the fact that they are batched (Matter et al., 2024), may impact responses. To ensure transparency, reviewers should make sure that authors 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 is 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 due to 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, reviewers must ensure authors transparently report and, ideally, justify their use of critical LLM parameters and application strategies such as batching, as these choices can significantly affect the study’s outcomes.

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. 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 where errors arise, authors must develop a process for handling them. Overall, reviewers should require that authors are transparent about post processing and error handling. For 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 post-processing and error handling prior to processing their data and make their plans public via pre-registration. Given the complexity of these tasks, however, some room should be allowed for plans to change. Reviewers should pay attention to whether the particular strategies used could skew the outputs towards desired inferences post-hoc. If authors, for example, validate their post-processed 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, reviewers should ensure authors clearly describe how they process data and handle errors in LLM outputs, including any post-processing 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. In LLM research, where small "tweaks" can heavily influence outcomes, preregistration is even more crucial. Researchers should preregister details such as the choice of LLM, model version and settings, prompts, processing of responses, including plans for handling ambiguous or outlier responses. This transparency facilitates unbiased conclusions, and it promotes replication by others, thus addressing concerns relating to validity and reproducibility in LLM research. Given the rapid evolution of LLM technologies, preregistration makes it easier for other researchers to understand, replicate, and build upon other researchers' work. This fosters credibility and aids scientific progress in analyzing human data with LLMs. In summary, reviewers ideally emphasize the importance of preregistration, in LLM research to facilitate transparency, validity, and reliability. This approach is crucial given the sensitivity of outcomes to specific LLM configurations and processing methods. Preregistration should contain validation and reliability strategies to minimize post-hoc choices that skew results toward desired outcomes.

Special Considerations for Simulating Human Data

If LLMs are used to simulate human data (i.e., produce primary data), several additional considerations are necessary. First, authors should validate, in the context of the research question, that the LLM generates data in a human-like manner. Ideally, reviewers should recommend that authors repeat a subset of the tasks with human participants to test whether the model and humans align. If this cannot be done (e.g., due to the sensitive nature of the tasks), then authors should ensure that the models show the capabilities that humans would use to produce the data. For example, if humans would solve a task by inferring others' mental states, the authors may show that the model has Theory of Mind capabilities. Alternatively, authors may impose theory-based constraints on the model (e.g., its output choices) to align it with human behavior based on past findings. While authors can, and often do, reference past works that demonstrate LLMs' social or cognitive

Table 3

Reviewer guidelines to ensure robust LLM findings.

Potential Issues	**Strategy**	**Details**
Unfounded / coincidental outputs	Confirm LLM outputs with human data	Reviewers should ensure 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	Ensure authors report and rationalize all LLM settings, such as "temperature," to understand their impact on the model's output variability and decision-making process.
Data processing bias	Transparent data processing and error handling	Reviewers should verify that authors disclose their data processing methodologies, including how they handle unexpected or outlier LLM outputs, to prevent biasing results towards desired outcomes.
Data dredging in LLM outputs / Selective reporting	Preregistration	Authors should 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".

capabilities relevant to the task, reviewers should exercise caution. As noted, when discussing model updates, the time elapsed since model development is crucial. Evolving models can have changing performance and capabilities. Therefore, reviewers may advise using stable models with established capabilities confirmed by prior research, such as open-source models that have publicly available model weights.

Second, authors need to consider the kind of data being simulated. For example, some research prompts LLMs to solve a set of tasks and then reports the (average) performance to infer general LLM capabilities or prompts an LLM to respond to a set of stimuli and then compares the responses to human samples (Bang et al., 2023; Coda-Forno et al., 2023; Dillion et al., 2023; Horton, 2023; H. Liu et al., 2023). These methods essentially simulate average outcomes rather than diverse individual responses, which can miss important nuances such as those across demographics. 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 which should be accounted for and justified (e.g., in pre-registrations). Additionally, 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 within populations (Beck et al., 2024; Durmus et al., 2023; Santurkar et al., 2023). Thus, reviewers should recommend justifying data generation methods and how they ensure the simulated data reflects the target population.

Lastly, 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 shows 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. For example, by creating completely novel stimuli or by applying "unlearning" techniques, which aim to remove the impact of “target data”, such as copy-righted material, psychological tests, or benchmark datasets, 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” which 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, e.g., 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 that are only accessed via API or when fine-tuning is not feasible.

Conclusion

LLMs are powerful tools that can automate previously time-consuming or expensive tasks. They have the potential to expand the scope of behavioral and psychological research by allowing for the analysis of larger qualitative datasets 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 some key areas for reviewers to attend to, focused on reliability and validity of LLM generated data. It is important to keep in mind that this technology is advancing rapidly, and best practices are likely to change in the future. As a result, providing detailed information like computer code, along with thorough explanations of methods, is extremely important as they will allow future generations of researchers to understand, and, if necessary, revise, research conducted with the current generation of LLMs. Challenging as it may be, the best reviewers will keep abreast of the latest guidance for reviewing LLM based behavioral and psychological research. As with any new technology, LLMs open up new horizons while generating new pitfalls, and this guide aims to help researchers who study human behavior and psychology benefit from this technology while avoiding some of the methodological challenges they pose.

Author Contributions

S.A: Conceptualization, Methodology, Investigation, Writing - Original Draft, Visualization

A. S. Z.: Conceptualization, Methodology, Investigation

A. M.: Conceptualization, Writing - Original Draft

D. B.: Conceptualization

M. D.: Conceptualization, Supervision

Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2023). Perils and opportunities in using large language models in psychological research. *PsyArXiv. URL: https://osf.io/preprints/psyarxiv/d695y*.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint 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. *International Conference on Machine Learning*, 337–371.

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 preprint 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.

Atari, M., Omrani, A., & Dehghani, M. (2023). Contextualized construct representation: Leveraging psychometric scales to advance theory-driven text analysis.

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

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. https://arxiv.org/abs/2302.04023

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chen, J., & Yang, D. (2023). Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*.

Chen, L., Zaharia, M., & Zou, J. (2023). Analyzing chatgpt's behavior shifts over time. *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Chiang, C.-H., & Lee, H.-y. (2023). Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

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 preprint arXiv:2304.11111*.

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

Davis, J., Van Bulck, L., Durieux, B. N., Lindvall, C., et al. (2024). The temperature feature of chatgpt: Modifying creativity for clinical research. *JMIR Human Factors*, *11*(1), e53559.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., et al. (2023). Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Eldan, R., & Russinovich, M. (2023). Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

Ericsson, K. A., & Moxley, J. H. (2019). Thinking aloud during superior performance on tasks involving decision making 1. In *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*.

Fujita, H., et al. (2022). Prompt sensitivity of language model for solving programming problems. *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)*, *355*, 346.

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.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

He, J., Wallis, F., Gvirtz, A., & Rathje, S. (2024). Artificial intelligence chatbots mimic human collective behaviour.

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

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, *8*(1), 216–225.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ke, L., Tong, S., Chen, P., & Peng, K. (2024). Exploring the frontiers of llms in psychological applications: A comprehensive review. *arXiv preprint arXiv:2401.01519*.

Kennedy, B., Jin, X., Davani, A. M., Dehghani, M., & Ren, X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*.

Kennedy, B., Reimer, N. K., & Dehghani, M. (2021). Explaining explainability: Interpretable machine learning for the behavioral sciences.

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2023). Bloom: A 176b-parameter open-access multilingual language model.

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., et al. (2024). Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Liu, Z., Dou, G., Tan, Z., Tian, Y., & Jiang, M. (2024). Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058.*

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2021). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786.*

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., & Kolter, J. Z. (2024). Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121.*

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

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. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.

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

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, *135.*

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2023). Gpt is an effective tool for multilingual psychological text analysis.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *International Conference on Machine Learning*, 29971–30004.

Serrano, S., & Smith, N. A. (2019). Is attention interpretable? *arXiv preprint arXiv:1906.03731.*

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.*

Tabone, W., & de Winter, J. (2023). Using chatgpt for human–computer interaction research: A primer. *Royal Society Open Science*, *10*(9), 231053.

Thaker, P., Maurya, Y., & Smith, V. (2024). Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*.

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

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 preprint arXiv:2307.03941*.

Zhao, Y., Zhang, R., Li, W., Huang, D., Guo, J., Peng, S., Hao, Y., Wen, Y., Hu, X., Du, Z., et al. (2024). Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 1–55.

Appendix

Checklist For Reviewing LLM Papers

- [**Optional**] Check if the study is pre-registered.
  - Are the methods, including models, parameters, and validation strategies, registered?
  - Does the preregistration allow for a full understanding of the intended experimental design, data analysis plan, and how results will be interpreted?
  - How closely does the study follow the preregistered protocols, and are any changes justified with transparent reasoning?
- [**Required**] Check if the model is stable (i.e., does not change over time) and accessible (i.e., can be used for replication).
  - If the model is not stable:
    - Check if an acceptable reason is provided.
    - Check if the model's exact name and query date are provided (to track model changes).
    - Check if the limitations of choosing the specific model, at query date, are provided.
    - [**Optional**] Ask to validate with a stable model.
  - If the model is not accessible:
    - Check if an acceptable reason (exclusion of accessible models) is provided.
    - [**Optional**] Ask to repeat with an accessible model.
- [**Required**] Check if all materials for replication are available:
  - Code
  - Model parameters and settings such as temperature.

  - Prompts
  - Data for fine-tuning
  - Any other study material (e.g., questionnaires, human validation data)

- [**Optional**] Check if the code runs without errors.
  - Check if replicated results align with reported results by a reasonable margin.
  - Evaluate strategies to account for LLM randomness.
    - Aggregation type (e.g., means and standard deviations of repeated runs).
    - Justification for aggregation (or lack thereof).
- [**Required**] Check if the reported results and inferences are justified.
  - Were the LLM outputs validated with human data or other justifiable data?
    - Is the achieved accuracy sufficient?
    - Is the achieved accuracy discussed (e.g., comparison to other methods)?
  - 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?
    - If not, is the applied prompt strategy and the model settings justified?
  - Evaluate data processing and error handling.
    - Is the data processing and error handling reasonable (e.g., justified outliers)?
    - Is the data processing and error handling biased toward the desired outcomes?