NBER WORKING PAPER SERIES

12 BEST PRACTICES FOR LEVERAGING GENERATIVE AI IN EXPERIMENTAL RESEARCH

Samuel Chang
Andrew Kennedy
Aaron Leonard
John A. List

Working Paper 33025
http://www.nber.org/papers/w33025

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2024

We appreciate insightful comments from Kyle Boutilier, Brian Jabarian, Alex Kim, and Connor Murphy. All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w33025

12 Best Practices for Leveraging Generative AI in Experimental Research
Samuel Chang, Andrew Kennedy, Aaron Leonard, and John A. List
NBER Working Paper No. 33025
October 2024
JEL No. C9, C90, C91, C92, C93

ABSTRACT

We provide twelve best practices and discuss how each practice can help researchers accurately, credibly, and ethically use Generative AI (GenAI) to enhance experimental research. We split the twelve practices into four areas. First, in the pre-treatment stage, we discuss how GenAI can aid in pre-registration procedures, data privacy concerns, and ethical considerations specific to GenAI usage. Second, in the design and implementation stage, we focus on GenAI's role in identifying new channels of variation, piloting and documentation, and upholding the four exclusion restrictions. Third, in the analysis stage, we explore how prompting and training set bias can impact results as well as necessary steps to ensure replicability. Finally, we discuss forward-looking best practices that are likely to gain importance as GenAI evolves.

Samuel Chang
University of Chicago
Booth School of Business
5807 S. Woodlawn Ave
Chicago, IL 60637
United States
samuel.chang@chicagobooth.edu

Andrew Kennedy
University of Chicago
andrewkennedy@uchicago.edu

Aaron Leonard
University of Chicago
aaronleonard@uchicago.edu

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and Australian National University
and also NBER
jlist@uchicago.edu

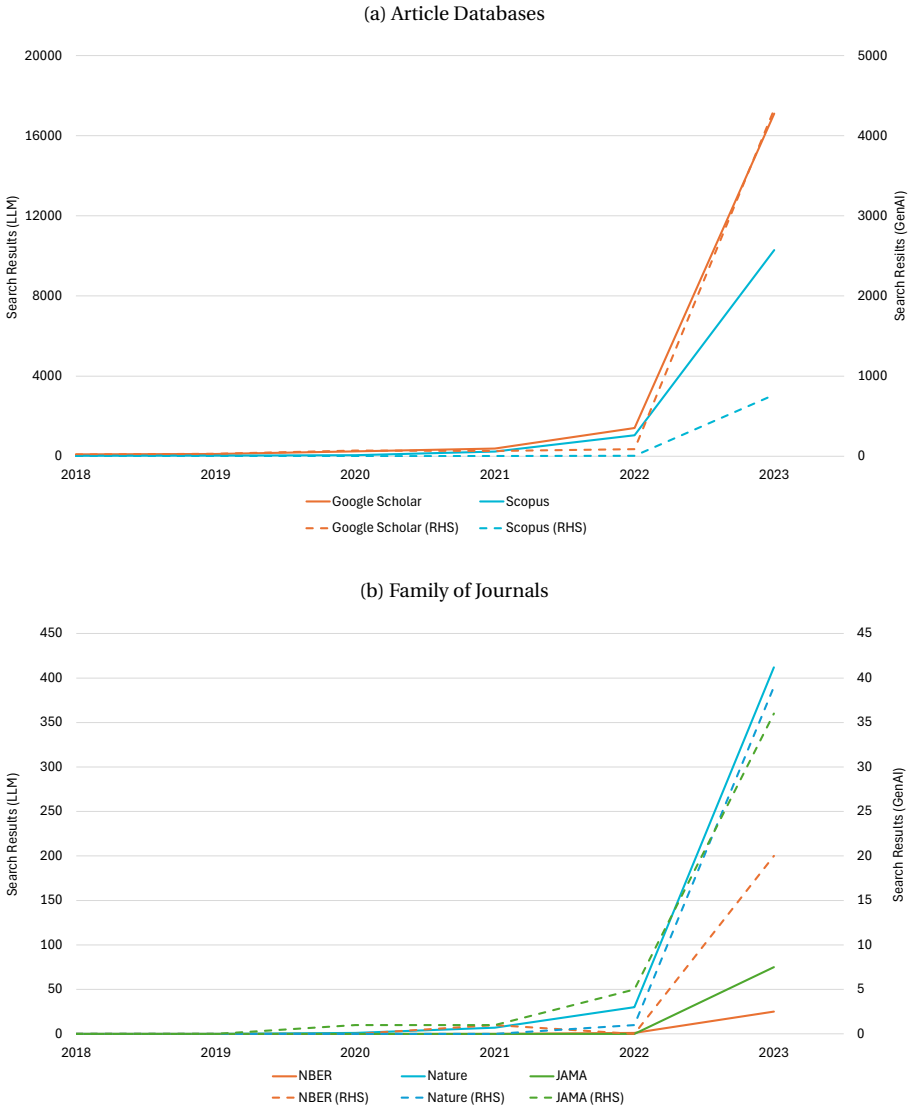
1. Introduction

Generative AI (GenAI)—artificial intelligence that can generate text, images, audio, and data—not only has the potential to deeply disrupt the workplace (Brynjolfsson et al., 2023) and important governmental sectors (Wachter & Brynjolfsson, 2024), it also holds promise in opening new avenues of scientific exploration and lowering the marginal cost of doing research. This is especially true in experimental settings, where design and data generation are vital aspects of the research process. GenAI, for instance, can aid in designing survey questions and treatments, conducting qualitative interviews (Chopra & Haaland, 2023), classifying open-ended responses, generating code, analyzing high-dimensional data, and even generating novel hypotheses (Ludwig & Mullainathan, 2024). The inclusion of GenAI within the research process therefore pushes the frontier through novel approaches to pre-treatment procedures, design, implementation, and analysis while simultaneously enabling researchers to more effectively fulfill internal and external validity in an ethically responsible manner.

At the same time, the lower marginal cost of doing research introduces a dimension of moral hazard: researchers may naively utilize GenAI without fully understanding the additional pitfalls and risks associated with its use. These include operational risks such as training set suitability and hallucinations, ethical risks relating to data security or model-training biases, and behavioral risks relating to confirmation bias or priming. Interest in GenAI, moreover, has increased exponentially over the last couple of years, expanding the population of researchers eager to reap the benefits of GenAI but whom are also equally as prone to the risks.

The top panel of Figure 1 shows the number of academic outputs referring to GenAI or large language models (LLMs) in both Google Scholar and Scopus. The bottom panel of Figure 1 shows its usage in various families of journals: Economics (NBER), Nature, and JAMA. The explosion of GenAI and LLMs is impressive and occurs across many scientific fields and journals. In our training, such a rampant increase has few parallels and mirrors observed price increases one finds in asset bubbles, such as the Beanie Baby craze in the 1990s, though we suspect that the GenAI movement is here to stay rather than a transitory phenomenon.

Figure 1: Number of Academic Outputs Referring to GenAI or LLMs



Currently, only a few studies outline risks and use cases associated with GenAI in experiments (Charness et al., 2023; Korinek, 2023; Bail, 2024), but, to the best of our knowledge, there does not exist a systematic framework to navigate these pitfalls while realizing the benefits of GenAI and maintaining the highest ethical standards. This paper establishes such a framework of twelve best practices, though these practices are neither exhaustive nor permanent as GenAI tools and applications continue to develop at an extraordinary pace. Rather, they represent what we believe at present to be important research considerations given current tools and use cases.

The remainder of the paper proceeds as follows. Section 2 discusses how GenAI con-

tributes to or detracts from the twin overarching experimental goals. Section 3 focuses on the pre-treatment stage, discussing preregistration, data privacy, and ethics. Section 4 focuses on the design and implementation stage, exploring how GenAI influences mechanism discovery, documentation and piloting, and fulfillment of the exclusion restrictions. Section 5 focuses on the analysis stage, examining prompting, training sets, and replicability.Section 6 highlights forward-looking considerations that we foresee as increasing in relevance as GenAI develops in sophistication. Section 7 concludes.

2. Twin Experimental Goals

Since the Renaissance, laboratory experiments have been a cornerstone of the scientific method. Increasingly, researchers in the social sciences have turned to the experimental model of the physical sciences as a method to understand human behavior. A key underpinning within the social science movement that deviates from the physical sciences is the use of randomization, which gained prominence nearly a century ago with the foundational work of Fisher, Neyman, and others.

By now, experimentation has expanded to occupy nearly every corner of the social sciences, with scholars scrutinizing how behavioral models explain behavior within communities, Walmart shoppers, day laborers, and Chief Executive Officers. For our purposes, List (2024a) provides definitions of the two overarching experimental goals:

> ***Experimental Problem 1 (EP1):*** *Measuring the causal impact of treatments and determining relevant mediators and moderators in an ethically responsible manner.*[1]
> ***Experimental Problem 2 (EP2):*** *Predicting whether the causal impacts of treatments implemented in one environment transfer to other environments, be them spatially, temporally, or scale differentiated.*

EP1 is the first bridge an experimenter must cross to construct a successful experiment. The first part of EP1 is understanding the "effects of causes," or what the literature denotes as internal validity. The second part of EP1 demands learning the "causes of effects," compelling an analysis of mediators and moderators. While the core of research relates to EP1, since therein lies important tests of theory and enhanced understanding of the world, our explorations in many cases should also be designed with an eye toward generalizing, as in EP2.

EP2 speaks to external validity but also encapsulates a broader set of experimental goals. The external validity problem relates to whether we can transport insights gained from one

[1]Mediators are channels through which treatments impact outcomes. Moderators are covariates that influence treatment effect sizes.

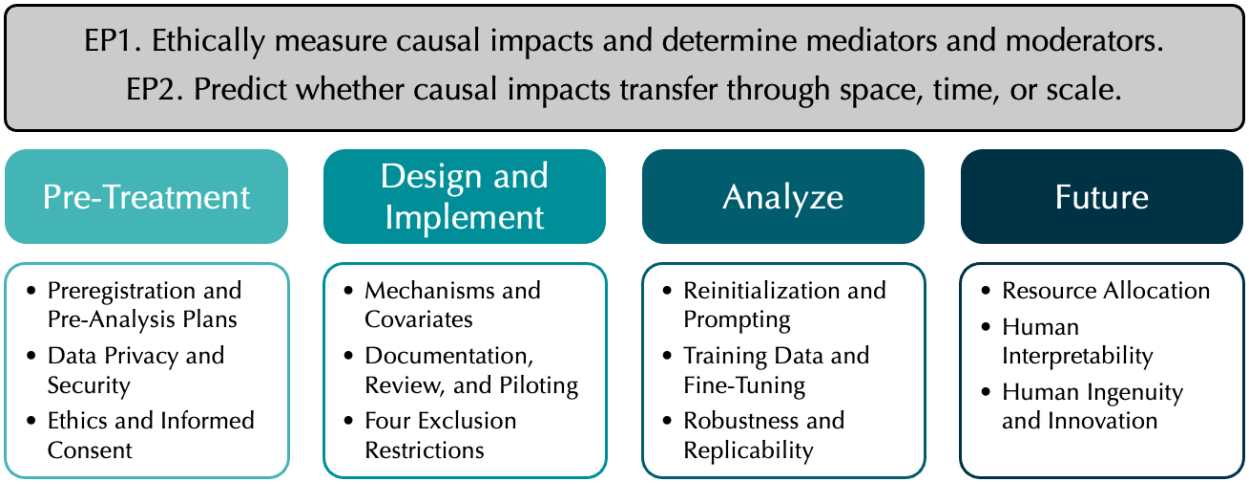


Figure 2: Using Generative AI to Fulfill Experimental Goals

population or domain to another, be that through space, time, or scale. EP2 is often the ultimate goal of social science experiments because, though internal validity can establish truths, the applicability of the result to distinct and larger samples is often of great importance, especially in studies exploring the efficacy of policy interventions or health initiatives. Most of the studies summarized in Figure 1 include a contribution within EP1 or EP2.

Demonstrating that EP1 and EP2 are fundamental to our discussion, Figure 2 projects EP1 and EP2 over all of our best practices. GenAI has both the potential to contribute to and detract from the fulfillment of our two experimental goals across the entire experimental spectrum: from pre-treatment chores to data analysis to replication. Each of our best practices for using GenAI in experimentation is motivated from EP1, EP2, or both, and each recommendation seeks to maximize internal and external validity through the careful usage of GenAI tools. In this way, our best practices remain focused on the essential components of an effective experiment.

3. Pre-Treatment Stage

In this section, we outline three best practices for researchers when using GenAI in the pre-treatment stage, related to alleviating the administrative burden of pre-treatment procedures, the privacy of data, and ethical considerations.

3.1. ***Best Practice 1.*** *Use GenAI to lower the administrative burden of pre-treatment procedures without lowering quality.*

As academic journals increasingly require pre-registration and pre-analysis plans (PAPs), the time and effort associated with these tasks are rising and could become prohibitively costly for researchers with fewer resources. Despite differing views on the necessary level of detail for

pre-treatment procedures, GenAI offers an opportunity to standardize and democratize this process for all researchers by reducing the resources required for these tasks in addition to harmonizing the formatting, organization, and content of pre-treatment paperwork. While we hope this can address the significant variation in pre-registration and PAP quality that currently exists, there is risk that researchers become overreliant on GenAI thereby leading to diminished quality.

To maintain a high quality while lowering the resource investment of pre-treatment procedures, we suggest that researchers use the following workflow that can be adapted for various projects as needed. First, upload a corpus of high-quality pre-treatment documents from existing projects to train the model. Second, upload the unformatted survey instrument(s), experimental instructions, and any other details necessary to include in the pre-treatment paperwork. Third, instruct GenAI to create draft pre-treatment documents based on the examples and unformatted information. Fourth, review and iterate on the GenAI output, minimizing the risk of over-reliance on model output and making sure to thoroughly utilize human expertise. Throughout the process, researchers should remain actively engaged to ensure that the information included is relevant so that the future implementation and analysis process is efficient and effective.

3.2. ***Best Practice 2.*** *Disallow public model training on inputted data and ensure that AI platforms have adequate data privacy measures.*

The usage of GenAI brings with it unprecedented data privacy considerations: if a public GenAI platform is used to analyze survey responses, researchers do not know where the data is stored, if appropriate steps are taken to protect the data, or whether the data will be used to train a future model which may unintentionally disclose inputted data to anyone using the model.[2] While most GenAI platforms publicly disclose their privacy policies, these are often broad and fail to reach the specificity researchers should require to satisfy themselves and ethics review committees. For instance, OpenAI, the platform behind ChatGPT, states in their data security and privacy policies that (1) "OpenAI may securely retain API inputs and outputs at varying lengths," (2) "ChatGPT is trained based on ... information that our users or our human trainers provide," and (3) "when you use our services for individuals such as ChatGPT or DALL•E, we may use your content to train our models." Such notions clearly compromise the responsibility of researchers to maintain clear oversight of collected data and ensure that it is not being abused.

[2] See https://www.aporia.com/learn/understanding-the-threat-of-data-leakage-in-genai-apps/ for a discussion of data leakage concerns based on a report by Gartner Research.

To maximize data security and autonomy when using public GenAI models, we recommend researchers take one of two steps. First and foremost, researchers should familiarize themselves with several platforms' data security and privacy policies and prioritize selecting platforms that do not collect or store user-inputted data for future model training. Second, if researchers must use a platform that collects user-inputted data for training, they should follow platform procedures to attempt to disallow model training from their data, such as by submitting a privacy request. While this does not address the security concerns of data storage as in step one, it alleviates concerns of accidental data leakage to other platform users. If neither of these steps is feasible, researchers may also consider fine-tuning a GenAI model which guarantees autonomy over data security and privacy, albeit requiring a significant investment of time, effort, and financial resources (Dell, 2024).

3.3. ***Best Practice 3.*** *Carefully consider additional risks to participants and elicit informed consent for GenAI use to ensure the highest ethical standards.*

GenAI changes the dynamic between research objectives and participant welfare. The speed at which GenAI is evolving exposes participants to potential harm that researchers do not yet fully understand and account for in their designs. At the same time, experimental policies implemented by ethics review bodies lag the potential risks associated with the fast-paced development of new GenAI methods. To uphold ethical research practices in the face of this disparity, we believe that a greater burden must fall on researchers, at least in the short term, to scrutinize the risks of GenAI within their own studies.

Specifically, we recommend that researchers proactively and critically assess their use of GenAI against the existing ethical principles of their jurisdiction to identify any possible shortcomings. In particular, these considerations are likely distinct across different research cultures which place different weights on the many dimensions of ethics. For instance, some ethics committees may find the additional risks of GenAI in a given project sufficient to reject the proposal on the grounds of potential undue harm to participants while ethics committees in other cultures may place lesser weight on these risks relative to the potential insights gained and approve the proposal. When working with researchers or participants from different cultures, we thus recommend that researchers consider foreign acceptance of GenAI usage and comply with foreign ethical norms insofar as these norms provide higher ethical standards than the researchers' domestic region. In this manner, we advise a "race to the top" rather than a "race to the bottom" or "weakest link" approach to ethics committee approval.

Informed consent is also more complex within the uncharted territory of GenAI use in experiments. This can be split into two dimensions: (1) participants' interactions with GenAI during the experiment, and (2) participants' preferences over the handling of their data. With

respect to (1), GenAI can be potentially harmful to participants because it can adapt to conform to participants' hostile or discriminatory behaviors and hallucinate to present false information as fact (Dev et al., 2021). Both of these risks are exacerbated by the inability of participants to differentiate whether they are interacting with humans or GenAI, which may lead participants to update their real-world beliefs under the false impression that they interacted with a human. With respect to (2), participants may, for a variety of reasons including cultural norms and privacy concerns, be willing to have their data analyzed by people but not GenAI. Taken together, we strongly recommend researchers seek explicit consent from participants to use GenAI within the experiment and/or in analyzing the collected data.[3] In the spirit of informed consent, researchers should clearly disclose in a manner accessible to the participant when GenAI will be used and provide a contextualized outline of the potential risks stemming from GenAI interactions and/or analysis before freely seeking consent from participants. If consent is not given for either interaction or analysis, irrespective of the reason, researchers must respect these preferences and not include the participant from the experiment.

4. Design and Implementation Stage

In this section, we outline three best practices for using GenAI in the design and implementation stage, specifically related to: identifying mediators, moderators, and scalable opportunities; piloting, documenting, and reviewing AI inputs and outputs; and the exclusion restrictions necessary to achieve internal validity.

4.1. ***Best Practice 4.*** *Use GenAI to identify new mediators and moderators while expanding the original design using Option C Thinking to optimally generate evidence that scales.*

Although GenAI opens up new opportunities to identify causal channels, heterogeneity, and scaling challenges, there is risk that these opportunities remain unrealized. Such risks can be decomposed into two different margins. First, the identification of potential mediators, moderators, and how ideas scale is at present an arbitrary and inexhaustive process, with the larger space of testable hypotheses exacerbating the paradox of choice and increasing the likelihood that important aspects are overlooked. Second, there is a disjoint between the novel mediators, moderators, and new treatment arms feasible via GenAI and traditional parsimonious theory, resulting in key information being neglected due to selective inattention.

To help focus research designs on a narrower set of relevant hypotheses, we recommend researchers utilize another aspect of GenAI—specifically, that GenAI mimics latent information

[3] The exception to this are studies where deception is required to test the desired hypothesis, such as an evaluation of whether people can distinguish between AI-generated content or human-generated content.

about human behavior that can be cheaply and systematically explored by researchers when designing an experiment. This can be as simple as asking LLMs for ideas of mediators and moderators given a research context to more complicated exercises, such as using GenAI to generate synthetic participants with latent human characteristics and simulating experiments to elicit relevant mediators and moderators (Horton, 2023). Within this area, we strongly encourage researchers to provide GenAI with a general structural framework as the basis to organize and communicate ideas, which has recently been shown to dramatically improve model predictions in the case of structural causal models (Manning et al., 2024).

In addition, this use of GenAI lends itself to considering external validity and how ideas scale (Banerjee et al. (2017)). Given an existing experimental design, for instance, researchers can test an identical design on a synthetic sample to generate suggestive evidence on whether similar conclusions may hold in different settings, enabling researchers to quickly assess the populations and situations that the collected results can transfer to and whether they should reconsider their own design. Likewise, such explorations can examine how the benefit-cost profile evolves at scale and whether interventions will experience a voltage drop as early as the design phase (List, 2022). In this spirit, GenAI can play an important role in creating treatment arms that provide the necessary insights and produce the type of policy-based evidence that the science of scaling demands from the beginning. This is denoted "Option C Thinking" (List, 2024b), and currently its most demanding aspect is the generation of relevant ideas and treatment arms to provide a good test of scaling, a task well suited for GenAI. We do caution, however, that GenAI in no way recuses researchers from responsibility over the research design, as researcher discretion must always be exercised to determine which potential channels or sources of heterogeneity are relevant and which are trivial or spurious.

4.2. ***Best Practice 5.*** *Never delegate data collection or experimental procedures to GenAI without appropriately piloting, documenting, and reviewing its roles and output.*

GenAI does not reduce the responsibility of researchers to carefully vet their own studies. On the contrary, GenAI makes piloting, documentation, and review even more important, as the researcher must gain awareness of unforeseen concerns stemming from GenAI involvement and correct any unexpected inaccuracies from GenAI output.

During piloting, we provide three examples of steps researchers should take to ensure the integrity of GenAI use. First, if researchers use GenAI models to construct survey questions, they should vet the questions using a similar process as if they had written them from scratch. This is consistent with the general principle that all GenAI outputs should be subject to human review for quality control. Second, if researchers use GenAI models to identify possible mediators, moderators, and Option C arms, they should test these for plausibility during piloting to

ensure that any irrelevant or uninterpretable variables do not remain part of the research design. Third, researchers should be highly alert to ethical concerns that arise during pilot testing. Even if the ethics committee initially approves a study, GenAI involvement may bring with it novel ethical considerations that may not become apparent until pilot tests are performed.

Even after correcting issues identified during piloting, researchers should maintain a detailed record of the role GenAI plays and the output it produces throughout each step of the implementation process. Specifically, there should be no intermediate steps involving GenAI that are observable at the time of interaction but not precisely documented. Further, researchers should critically review the documentation both during the experiment, to dynamically respond to detected errors, and after its conclusion, to ensure that no mechanical errors remain. These steps are consistent with the principle of human review and necessary to assess the overall procedural integrity and maintain maximum replicability of the experimental results.

4.3. ***Best Practice 6.*** *Consider how GenAI both supports and challenges upholding the four exclusion restrictions necessary for internally valid experiments.*

At the backbone of the experimental paradigm lies the four exclusion restrictions: (1) observability; (2) complete compliance; (3) statistical independence; and (4) the stable unit treatment values assumption (SUTVA) (List, 2024a). Violating any of these restrictions compromises the validity of an experiment—including by biasing estimated effect sizes or creating sample selection problems—making it critical that researchers aptly design their experiments to uphold the exclusion restrictions and correctly identify constructs of interest. GenAI provides a number of means to bolster this process, such as through standardizing interventions across participants or minimizing non-random attrition by providing low-cost channels for follow-up, but is equally a threat to the exclusion restrictions, such as through introducing unobservable heterogeneity to interventions or non-random attrition from differential response rates to GenAI follow-up, respectively.

We recommend researchers adopt a risk-first, opportunity-second framework to determine how to leverage GenAI in support of the exclusion restrictions. This framework is conceptually simple: first, consider what steps, if any, can be taken to minimize threats to the exclusion restrictions introduced by GenAI without decreasing its effectiveness; second, if stage one is insufficient in preserving the exclusion restrictions, consider whether the application of GenAI can be adjusted such that the threat becomes insignificant. Rather than outlining a complete set of recommendations for preserving the exclusion restrictions using GenAI in every context—which is infeasible as every experiment has its own idiosyncratic threats—we instead limit our discussion to the context of GenAI chatbots and SUTVA to serve as an example of how researchers could implement the risk-first, opportunity-second framework.

One part of SUTVA is the no hidden treatments assumption, which stipulates that participants within the same treatment arm receive an identical intervention. GenAI can help safeguard this assumption by ensuring that interventions are uniformly administered. For instance, the characteristics of human enumerators are widely documented as having significant impacts on outcomes, especially those related to religion, ethnicity, gender, and politics (Di Maio & Fiala, 2020); using GenAI chatbots in place of human enumerators in these contexts provides a means of eliminating heterogeneity in the intervention. GenAI chatbots, however, can adapt to the perceived preferences and views of participants through repeated interactions, which presents a threat to SUTVA if such adaptations vary systematically based on participant characteristics. Applying the risk-first, opportunity-second framework to resolve this quandary, researchers should first take steps to mitigate any systematic adaptation based on participant characteristics without decreasing capacity of the GenAI chatbot, such as providing chatbots with a set of guidelines and common instructions. Assuming for demonstration this action proves insufficient, researchers should then consider steps which may decrease the capabilities of GenAI as enumerators, such as lowering the temperature (creativity) of the GenAI chatbots to limit their variability across participants. By addressing the problem in such a two-stage approach, we ensure that researchers mitigate the risk that chatbots themselves cause hidden treatments while maximizing the standardization benefits of using GenAI chatbots as enumerators, thus upholding SUTVA and the integrity of the experimental findings.

5. Data Analysis Stage

At the data analysis stage, GenAI can support researchers to clean data, analyze high-dimensional data and unstructured responses, use baseline information to enhance precision, and surmise narratives (Charness et al., 2023). This section provides best practices for specifying prompts, selecting an appropriate training set, and maximizing the replicability of findings given that GenAI models do not output identical results each time they are run.

5.1. ***Best Practice 7.*** *Ensure consistent reinitialization and provide prompts that capture the construct of interest.*

Much like humans, GenAI attempts to use all information available, including the context from previous prompts, to inform the response to a given prompt. As a result, GenAI models can be subject to priming biases when searching through data or unstructured survey responses, leading to GenAI generating conclusions from data based on prior interactions. This is especially relevant in scenarios where researchers are testing a particular hypothesis but prompt GenAI to highlight themes of the data. By previously revealing the hypothesis to

GenAI—such as by inquiring how to best phrase a question—GenAI is primed to search for related patterns in the data and therefore any identified themes are subject to a type of systematic confirmation bias (Perez et al., 2023).

To minimize this risk, we propose that researchers are cognizant of the ordering of prompts and reinitialize the GenAI model where appropriate to clean the 'memory' of GenAI to guarantee that information spillovers do not occur between prompts. In addition, we propose that, where possible, the wording of prompts is neutral in character but specific in direction to attenuate the likelihood that prompts are priming GenAI towards particular conclusions. As with most aspects of experimental design, different prompts can be tested during piloting to see which one captures the construct of interest most accurately (Si et al., 2023).

5.2. ***Best Practice 8.*** *Carefully consider training data and, when appropriate, manually train GenAI models used for analysis while simultaneously using ML to enhance experimental precision.*

Analyzing data using GenAI — such as using GenAI for sentiment analysis of social interactions or open-ended responses — requires the specification of appropriate training data for the model. Many researchers, however, utilize publicly-available models whose default training data is usually sourced from universes such as the internet. Although this in and of itself is not a disqualifying factor, it is undesirable for two reasons. First, researchers are unable to control what information the model takes into account when analyzing data, which is especially problematic if training data introduces biases relating to ethnicity, sexuality, gender, or religion (Grossmann et al., 2023; Santurkar et al., 2023). Second, the model may not be optimally trained to handle the particular inputs that the researcher provides, whether this is due to the type of input, language of the input, or social context (Latif & Zhai, 2024; Mayfield & Black, 2020). A LLM determining British respondent's attitude to alcohol consumption trained on default internet data, for instance, could erroneously code responses as being of negative sentiment in relation to the word 'pissed' which relates to being drunk in Britain but angry elsewhere, skewing the analysis in a way that is unobservable to the researcher.

For these reasons, we recommend that researchers investigate fine-tuning a GenAI model specifically for their analyses and always ensure that the training dataset is relevant to their research setting. While the latter is a minimum requirement and may involve a separate phase for collecting training data when relevant existing data is unavailable, the former requires a non-trivial investment of time, effort, and financial resources; an option which may not always be optimal and decision of which rests with the research team. We posit, however, that such decision frameworks prioritize the principles of model transparency, performance, and accessibility to appropriately capture the benefit-cost profile relative to off-the-shelf GenAI models. For

guides on how to train LLMs, we point the reader to Howard & Ruder (2018) and Jeong (2024)

The final part of Best Practice 8 concerns statistical analysis of experimental data. Including covariates in the regression framework has the benefit of increasing statistical power at the risk of potentially overfitting models to the data—a trade-off that machine-learning methods are well-suited to optimally balance. We therefore recommend that researchers leverage machine-learning (ML) methods at the analysis stage. This not only increases statistical power for a given sample size but also limits the scope for data-snooping and abstracts from overspecifying specifications due to behavioral fallacies relating to the sunk cost of collecting covariates. Within the toolbox of machine-learning methods, we specifically highlight the machine-learning approach of List et al. (2024), who solve for the efficient regression adjustment in a large class of adjustments and find considerable gains to statistical power using GenAI. More specifically, they show that using ML techniques over linear regression adjustment can allow researchers to attain similar levels of statistical power with 4-8 percent fewer observations.

5.3. ***Best Practice 9.*** *Maximize the replicability and reliability of results through documentation and stress testing, allowing the broader research community to confirm experimental findings.*

The use of GenAI increases the difficulty of replication given that GenAI models do not consistently produce identical outputs with identical inputs (Chen et al., 2023; Belz et al., 2023). Compounding this further, the advent of GenAI has increased the perceived likelihood that research is unreliable owing to GenAI expanding the toolbox of possible unethical practices. These include: (1) using GenAI to fabricate data and complex simulated responses that mechanically align with a desired experimental hypothesis ; (2) taking advantage of additional researcher discretion when prompting GenAI to manipulate data analysis or p-hack results (Kenthapadi et al., 2023); and (3) exploiting the 'black box' nature of algorithms to impede transparency or heighten ambiguity. [4]

We recommend three possible steps to maximize replication given these additional constraints. First, researchers should organize replication packages so that replicators can reproduce their results using their own GenAI tools. In particular, the training data, raw experimental data, prompts, and version of the GenAI used need to be provided to enable replication. While this may still lead to distinct results due to how the replicator's model is either trained or how it analyzes data, this approach reduces the number of dimensions from which inconsistency with the original findings could arise and consequently suggests that the findings, if not replicable, are sensitive to choices such as model selection, training, and/or prompting.

[4] P-hacking is the practice of running many statistical specifications corresponding to similar hypotheses and citing on those with desired significance levels.

Second, researchers should specify different sets of prompts or data training methods and use these as different measures to evaluate the robustness of the main specification results. For instance, researchers could vary the temperature (creativity) of the GenAI model and analyze how different measures created under these different temperatures change the results. Alternatively, for data coded numerically, a researcher could construct an index or take the mean of the different measures and use this in their primary analysis (Alizadeh et al., 2023). Third, researchers should comprehensively and precisely disclose their intended uses of GenAI in pre-treatment paperwork to engender confidence that flexibility during the analysis stage does not inflate the false positive rate. As GenAI continues to develop at a rapid pace, we anticipate that researchers may increasingly need to deviate from pre-registration during the implementation stage. In these situations, researchers should disclose and justify such deviations in the manuscript and perform robustness checks around the modifications.

6. Future Considerations

GenAI is advancing at an exponential rate, as is its usage as shown in Figure 1. We foresee that the capability, functions, and usability of GenAI in the coming years will meaningfully exceed that of the present. Likewise, while GenAI is presently an unfamiliar technology to the majority of the global population, we also expect that general aversion to using GenAI in research will abate as people become more accustomed to GenAI across many aspects of life. Owing to these two factors, we provide three best practices which, although less salient in the short term, we envisage to be nontrivial in the long term.

6.1. ***Best Practice 10.*** *Adapt resource allocation over different stages of experimental research as GenAI changes trade-offs.*

Realizing the benefits of GenAI augments a quintessential reality of experimental research: different elements of the research design are always in direct competition for resources, requiring researchers to trade off parts of the design to satisfy funding constraints. The use of GenAI alters these trade-offs, which we foresee as materially changing how resources are allocated across research components as GenAI becomes more capable. For instance, the use of GenAI to conduct interviews and record data not only reduces variation arising from different enumerators but significantly decreases the cost of hiring enumerators, which is usually a nontrivial allocation of the research budget in field experiments. These funds can then be allocated to other aspects of the research design such as piloting, increasing sample size to increase statistical power, increasing participation incentives, or changing the level of randomization. In contrast, the use of GenAI can also increase costs, the most significant of these generally being the time and funding associated with appropriating and training the GenAI model.

Given the rate of advance of GenAI, we recommend that researchers continually update their beliefs on the optimal allocation of resources across different stages of the research design. In particular, within this process, we recommend that researchers prioritize investment in elements of the research design which have a relative advantage in pushing the research frontier, including collecting additional covariates to determine effect channels or adding additional treatment arms to evaluate the impact of spillovers. The general governing process should be one of constantly comparing marginal benefits and marginal costs.

6.2. ***Best Practice 11.*** *Hypotheses generated through GenAI need to be interpretable to humans.*

GenAI has the power to notice patterns in data, a skill that previously was limited to the cognitive function of only a few intelligent species. What is more, GenAI can find patterns that humans cannot, including through processing vast quantities of high-dimensional data encompassing text and images in addition to traditional tabulated data. Given the increasing availability of such data, we expect that it will be common practice over the coming years for GenAI to produce numerous hypotheses from large inputted datasets. A limitation, however, is that such hypotheses may not be interpretable to humans, either because the hypotheses are themselves unexplainable as they are generated within a 'black box' or because they are not understandable as the language used to communicate the hypothesis is mechanical in nature (Doshi-Velez & Kim, 2017; Marcinkevičs & Vogt, 2020). When analyzing photo inputs, for instance, GenAI may hypothesize that 'if more than 40% of pixels on the lower half of the image are white, then there is a higher likelihood of the subject dying of heart disease.' This is neither explainable in that the GenAI cannot outline the logic underlying the hypothesis, nor is it understandable as it is not apparent what feature relates to 40% of pixels on the lower half of the image being white and its connection to heart disease.

We posit that GenAI-generated hypotheses must be tested against independent subjects to establish interpretability. The experimental approach is a natural analog to this process as, given a testable hypothesis, researchers have sufficient control over the environment such that they can experimentally manipulate mediating channels and moderating factors to identify a theory of change interpretable to humans. Alternatively, a recent approach by Ludwig & Mullainathan (2024) directly tests interpretability by prompting GenAI to generate pairs of synthetic images that vary only by a characteristic hypothesized by GenAI. Human subjects are then asked to identify the difference between images, with correct identification indicating that the hypothesis is interpretable and the failure to correctly identify the difference indicating that the hypothesis is not interpretable. We recommend that such approaches to interpretability are the baseline when using GenAI for hypothesis generation.

6.3. ***Best Practice 12.*** *GenAI is not a substitute for human ingenuity and innovation.*

Although GenAI is a powerful research tool, it should never be considered a substitute for human ingenuity. The rationale for this is twofold. First, it is possible that researchers unintentionally become dependent on GenAI, especially given the institutional pressures of producing research output. Since GenAI is presently trained on existing knowledge, it has certain difficulties developing new methodologies to shift the frontier of knowledge by evaluating hypotheses previously not testable, which we believe is core to the practice of research. Similarly, there is also risk that widespread reliance on GenAI leads to a homogenization of thought and methodology within disciplines that otherwise would have a robust variance of ideas. This selective inattention would then suppress different schools of thought within a discipline, stymieing knowledge creation against the spirit of Mill's free market of ideas (Mill, 1859). Second, there is risk that researchers remove themselves from the practicalities of the research process as GenAI becomes more sophisticated. This is a problem across all types of experiments but is especially relevant in field experiments where presence on the ground is key to diagnosing issues with randomization or the treatment itself, both of which would invalidate the experiment but go unnoticed when researchers are not present (Glennerster, 2017).

We therefore recommend that researchers partition time early on in the idea generation stage where GenAI is not utilized. This would ensure that researchers preserve their creative autonomy and pursue novel ideas that can shift paradigms by abstracting away from the existing body of knowledge. Moreover, we recommend that it is equally if not more important that researchers maintain an on-the-ground role in implementing experiments, including presence in the field to identify any breakdown in implementation and, in particular, any issues with the implementation of GenAI.

7. Conclusion

We present twelve best practices that provide a foundation for researchers to realize the benefits of GenAI while navigating the additional pitfalls and risks that GenAI simultaneously introduces. We do not claim that these best practices are exhaustive. Rather, they are important dimensions to consider for any researcher using GenAI in their research. Further, each of these twelve practices is flexible enough to enable researchers to push the frontier of knowledge through novel approaches to pre-treatment procedures, experimental design, data collection, and analysis but are sufficiently restrictive to mitigate the behavioral biases, implementation issues, and ethical concerns associated with using GenAI in research.

Although each best practice is relevant to all types of experiments, the exact level of relevance varies across disciplines and contexts. Best Practice 5 relating to piloting and documentation, for instance, is more relevant to field experiments where the loss from an implementa-

tion issue is greater than in lab experiments. Likewise, Best Practice 4 is more relevant to the social sciences where synthetic experiments can reveal insights about behavior but is not relevant to medical trials analysing physical outcomes. Determining the relative weight placed on these best practices is ultimately at the discretion of researchers given their specific context, and therefore is a subjective exercise prone to disagreement. We do believe, however, that the guidance provided through this framework best positions experimental research to be at the forefront of the research community in responding to the transformative impact of GenAI.

Of course, any such list will be incomplete and we suspect we, and other scholars, could create another dozen best practices. For example, while we discuss replications, integrity and reliability of results were largely left on the sidelines [5]. GenAI can be a powerful tool in curbing scientific fraud through several innovative approaches, including the following. (1) Data Integrity and Verification: GenAI can help verify the authenticity of research data by generating synthetic datasets to compare against reported results (Li et al., 2023; Kruschwitz & Schmidhuber, 2024). This can highlight inconsistencies or anomalies that may indicate fraudulent activity. (2) Automated Peer Review: GenAI can assist in the peer review process by identifying potential issues in research papers, such as data manipulation, plagiarism, or statistical errors (Ferrero et al., 2017). This can enhance the rigor and efficiency of the review process. (3) Behavioral Analysis: By analyzing patterns in researchers' behavior and writing styles, GenAI can detect unusual activities or deviations from typical practices that might suggest fraud (Braud & Søgaard, 2017). (4) Real-Time Monitoring: GenAI can be used to monitor ongoing research activities in real-time, providing alerts for any suspicious actions or deviations from standard protocols (Tagami et al., 2018). (5) Deepfake Detection: As deepfake technology becomes more sophisticated, GenAI can help detect and prevent the use of manipulated images, videos, or audio in scientific publications (Uchendu et al., 2024; Antoun et al., 2023). By leveraging these capabilities, GenAI can significantly enhance the integrity and reliability of scientific research, helping to prevent and detect fraudulent activities more effectively.

In the end, to deepen the stock of scientific knowledge, we are optimistic about the injection of GenAI into the experimental approach. This is due to two unique features of the experimental approach: selective data generation and successful replication. Perhaps the most exciting research in the coming decades will leverage GenAI to generate data across different settings, markets, subject pools, and moderators while ascertaining key mediation paths to increase our confidence in building the knowledge framework necessary for testing theories and providing empirical advice. Unlike yesterday's researcher who passively visited the pin fac-

[5] For one estimate of such misbehaviors, see (List et al., 2001), who report on various misbehaviors, from unethical research practices to classroom grading, of academic economists.

tory, tomorrow's researcher will surgically control the assignment mechanism, with the help of GenAI, to tinker within the pin factory. In turn, our scientific experiments will have a much better chance to drive real societal change.

References

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2023). Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks. *arXiv preprint arXiv:2307.02179*.

Antoun, W., Mouilleron, V., Sagot, B., & Seddah, D. (2023). Towards a robust detection of language model generated text: Is chatgpt that easy to detect? *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles*, 14–27.

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives*, *31*(4), 73–102.

Belz, A., Thomson, C., Reiter, E., & Mille, S. (2023). Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in nlp. In *Findings of the Association for Computational Linguistics: ACL 2023*, (pp. 3676–3687).

Braud, C. & Søgaard, A. (2017). Is writing style predictive of scientific fraud? *Proceedings of the Workshop on Stylistic Variation*, 37–42.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work. *National Bureau of Economic Research*.

Charness, G., Jabarian, B., & List, J. A. (2023). Generation next: Experimentation with AI. *National Bureau of Economic Research*.

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, *6*(2).

Chopra, F. & Haaland, I. (2023). Conducting qualitative interviews with AI.

Dell, M. (2024). Deep learning for economists. Technical report, National Bureau of Economic Research.

Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., et al. (2021). On measures of biases and harms in nlp. *Findings of the Association for Computational Linguistics*, 246–267.

Di Maio, M. & Fiala, N. (2020). Be wary of those who ask: a randomized experiment on the size and determinants of the enumerator effect. *The World Bank Economic Review*, *34*(3), 654–669.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ferrero, J., Agnes, F., Besacier, L., & Schwab, D. (2017). Usingword embedding for cross-language plagiarism detection. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 415–421.

Glennerster, R. (2017). The practicalities of running randomized evaluations: Partnerships, measurement, ethics, and transparency. In *Handbook of economic field experiments*, volume 1 (pp. 175–243). Elsevier.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108–1109.

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.

Howard, J. & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 328–339.

Jeong, C. (2024). Fine-tuning and utilization methods of domain-specific llms. *arXiv preprint arXiv:2401.02981*.

Kenthapadi, K., Lakkaraju, H., & Rajani, N. (2023). Generative AI meets responsible AI: Practical challenges and opportunities. *KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*.

Kruschwitz, U. & Schmidhuber, M. (2024). Llm-based synthetic datasets: Applications and limitations in toxicity detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying*, (pp. 37–51).

Latif, E. & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 100210.

Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10443–10461.

List, J. A. (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. Crown Currency.

List, J. A. (2024a). *Experimental Economics: Theory and Practice*. University of Chicago Press.

List, J. A. (2024b). Optimally generate policy-based evidence before scaling. *Nature*, *626*, 491–499.

List, J. A., Bailey, C., Patricia, E., & Thimas, M. (2001). Academic economists behaving badly? a survey on three areas of unethical behavior. *Economic Inquiry*, *39*, 162–170.

List, J. A., Muir, I., & Sun, G. (2024). Using machine learning for efficient flexible regression adjustment in economic experiments. *Econometric Reviews*, *44*(1), 1–39.

Ludwig, J. & Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, *139*(2), 751–827.

Manning, B. S., Zhu, K., & Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. *arXiv preprint arXiv:2404.11794v2*.

Marcinkevičs, R. & Vogt, J. E. (2020). Interpretability and explainability: A machine learning zoo minitour. *arXiv preprint arXiv:2012.01805*.

Mayfield, E. & Black, A. W. (2020). Should you fine-tune bert for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, (pp. 151–162).

Mill, J. S. (1859). *On Liberty*. Springer.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2023). Discovering language model behaviors with model-written evaluations. *Findings of the Association for Computational Linguistics*, 13387–13434.

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

Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., & Wang, L. (2023). Prompting gpt-3 to be reliable. *International Conference on Learning Representations*.

Tagami, T., Ouchi, H., Asano, H., Hanawa, K., Uchiyama, K., Suzuki, K., Inui, K., Komiya, A., Fujimura, A., Yanai, H., et al. (2018). Suspicious news detection using micro blog text. *32nd Pacific Asia Conference on Language, Information and Computation*, 648–657.

Uchendu, A., Venkatraman, S., Le, T., & Lee, D. (2024). Catch me if you gpt: Tutorial on deepfake texts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, (pp. 1–7).

Wachter, R. M. & Brynjolfsson, E. (2024). Will generative artificial intelligence deliver on its promise in health care? *Journal of the American Medical Association*, *331*(1), 65–69.