The Survey Statistician, 2025, Vol. 92, 30–44.

Successfully Navigating the Disruption AI will Bring to Survey Research

David M. Rothschild[1], **Trent D. Buskirk**[2], **Stephanie Eckman**[3], **D. Sunshine Hillygus**[4], **Frauke Kreuter**[3], **David Lazer**[5]

[1]Microsoft Research, USA, David@ResearchDMR.com, Corresponding Author
[2]Old Dominion University, USA, TBuskirk@odu.edu
[3]University of Maryland, USA, Steph@umd.edu / FKreuter@umd.edu
[4]Duke University, USA, Hillygus@duke.edu
[5]Northeastern, USA, D.Lazer@neu.ed

Abstract

Surveys are a core methodological tool in government, industry, and academia, providing essential data for theory development and evidence-based decision-making. As artificial intelligence continues its rapid advancement, it stands to fundamentally transform the entire survey lifecycle – from design and administration to analytics and reporting. Previous transitions to new technologies, such as telephone, internet, and non-probability surveys, led to divisions within the survey research community with real consequences for both the trajectory of research and trust in the industry. We believe the survey community should take proactive steps now to avoid similar challenges with AI integration. Specifically, our paper examines the potential benefits and risks AI introduces to survey methodology. We first identify promising research opportunities and innovations that merit further exploration. We then outline strategic recommendations for the survey research community to navigate this transition effectively, including guidelines for publication standards and research prioritization. Finally, we propose collaborative initiatives between AI specialists and survey researchers that could yield mutual advantages. This paper aims to influence the trajectory of research and implementation during this critical early phase. By addressing these considerations now, we hope to facilitate a more efficient and constructive integration of AI into survey research practices.

Keywords: survey methods, survey research, artificial intelligence (AI), large language models (LLM)

1 Introduction

Across academic fields, industry, and government, survey research plays a foundational role in generating data, informing theory, and shaping decision-making. As artificial intelligence (AI), with a strong focus on Large-Language-Models (LLM), continues to develop at an unprecedented pace, it is

Copyright © 2025, David M. Rothschild, Trent D. Buskirk, Stephanie Eckman, D. Sunshine Hillygus, Frauke Kreuter, David Lazer. Published by International Association of Survey Statisticians. This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

poised to radically transform how surveys are designed, administered, analyzed, and reported. Recent work concludes that survey research is among the occupations most exposed to automation by generative pre-trained transformer models (GPTs) (Eloundou et al. 2023). If past technological shifts are any guide, the integration of AI tools promises to improve efficiency, accessibility, and flexibility of survey design and implementation (Rothschild et al. 2024). But our history also offers a cautionary tale: when previous technological innovations such as the telephone, the internet, and non-probability panels were introduced, they fragmented the survey research community. Some embraced unquestioningly, while many others (especially within the existing survey research community) resisted reflexively. The result was a polarization of norms and practices within and across fields, leading to confusion among researchers and practitioners alike over what constituted rigorous, ethical, and effective survey work. As researchers apply AI tools throughout the survey lifecycle, the potential for similar fragmentation looms large particularly during this period of early and uneven adoption (Barrie, Palmer, and Spirling 2024).

Successfully integrating AI into survey research will require a collaborative effort to drive methodological innovation, while preserving fundamental values like transparency, accountability, and scientific rigor. This paper explores the key challenges and opportunities AI presents to survey research, providing a roadmap for both researchers and technologists. We examine areas where AI is poised to disrupt existing survey practices, highlighting the potential benefits, risks, and the critical research questions these disruptions raise. Our focus is on maintaining scientific rigor – particularly in areas such as data quality, representativeness, and bias – while also emphasizing transparency and accountability, including ethics and reproducibility. The goal is to safeguard the core principles of credible survey research, which could be tested or redefined in the AI era, while also preparing for the transformative impact of this technology. Drawing on lessons from the introduction of previous technologies, we outline strategic priorities and key actions to guide the field through this transition.

First, we must **monitor and evaluate AI use cases** across the survey lifecycle – design, administration, analysis, and reporting – documenting current applications and anticipating future ones that could improve data quality or reduce costs. This ongoing assessment will help clarify both the risks and rewards of AI integration and identify priority research gaps. While our review of existing use cases is necessarily incomplete in such a fast-moving field, it provides a starting place for this effort.

Beyond cataloging use cases, **evaluate AI models and tools** for their fitness for use, considering survey quality dimensions such as accuracy, reliability, timeliness, and comparability (Lyberg 2012). For example, AI-driven automation in data coding or imputation may improve timeliness, but must also be assessed for its impact on accuracy and consistency. Adoption decisions should be based not on novelty or convenience, but on rigorous evaluation aligned with the specific research context and quality standards. Academic and popular publications can help us create a forum for documenting how AI tools are being tested, used and incorporated within our field.

Perhaps a more critical precursor of publication is the need for the **development of evaluation frameworks and benchmarks** to determine if the use case actually works for the needs of stakeholders (even if, in isolation, they score well on key quality dimensions). What are the appropriate benchmarks for validating accuracy and how do they shift under various conditions and goals? What human in the loop standards are needed for various applications of LLMs? We are currently seeing widely varying yardsticks being used to evaluate the same applications. For example, recent studies have explored the use of large language models (LLMs) to generate synthetic survey responses, with some evaluations suggesting these responses are useful signals based on comparisons of means with existing data collections (Argyle et al. 2023; Masterton, Henriques, and Mclean 2025). However, other research demonstrates that while AI-generated responses may align with human data on average, they often diverge significantly on other critical indicators such as response distributions,

internal consistency, and subgroup variation – raising significant concerns about the utility of data (Bisbee et al. 2024; Heyde, Haensch, and Wenz 2025). The lack of ground-truth and potentially divergent use-cases could lead to very different evaluations of the same data.

The field's previous experience with internet surveys also highlights the importance of **transparency and disclosure**. Researchers must clearly document how AI tools are used throughout the survey process. Without such transparency, it becomes nearly impossible to assess the validity, replicability, or limitations of AI-mediated research. To incentivize openness, journals, conferences, and professional associations should adopt and enforce rigorous standards that make AI use auditable and replicable. This is no small challenge. At the most recent (May 2025) meeting of the American Association of Public Opinion Research (AAPOR), several presenters described their prompt engineering methods as proprietary, effectively shielding key methodological details from scrutiny. This highlights a growing tension: innovation in AI often occurs at the intersection of academia and industry, where commercial interests may conflict with scientific norms of openness and reproducibility. Many journals and associations need stronger policies requiring disclosure of commercial affiliations and funding. At the same time, collaborations and partnerships between survey statisticians, computer scientists, and behavioral researchers are critical. Ultimately, the responsible integration of AI into survey research will depend not only on technical innovation, but on a shared commitment to transparency, accountability, and cross-disciplinary collaboration.

We have learned the hard way that consensus among survey methodologists alone is not enough to establish effective standards. Given the diversity of stakeholders – including researchers and practitioners across industry, government, and academia – it is essential to **engage with adjacent fields** such as marketing, journalism, and the social sciences as well as those from the information technology and computer sciences and engineering fields. Involving these communities ensures that survey users actively participate in and benefit from evaluation efforts. Establishing shared benchmarks across disciplines will enable meaningful comparisons across studies, accelerate collective learning, and promote best practices. Survey researchers across fields must be equipped with the skills to critically assess and responsibly implement AI tools, while data consumers – including policymakers, journalists, and the public – need to be educated on how to interpret AI-mediated survey findings with appropriate context and caution.

Finally, we turn our attention to the AI and technology communities, highlighting ways they can more effectively engage with survey researchers to build tools that are not only technically impressive but also aligned with the ethical and methodological standards of the social sciences. We argue that **reciprocal collaboration between survey research and AI communities** is essential, not only to prevent harm, but to maximize the potential of AI as a force for good in evidence-based research.

2 Risks, rewards, and research for various potential disruptions

We consider use cases in four stages of the survey pipeline: design (translating research questions into a survey instrument and infrastructure), administration (fielding the survey), analytics (interpreting the data), and reporting (communicating findings to stakeholders). We consider the implications for data quality, drawing on the Total Survey Error (TSE) framework (Groves and Lyberg 2010), which identifies the following sources of survey error: specification error, measurement error, coverage error, nonresponse error, and processing error. We also consider the broader ethical and procedural implications, including respondent privacy, data security, transparency, and reproducibility. Based on the current state of the literature, we assess both the potential benefits and risks of integrating large language model (LLM) tools. We also identify key open research questions surrounding their use, emphasizing the need for continued investigation into their evolving impact on survey research.

2.1 Design Stage (AI as the Research Colleague): High Potential, Manageable Risk

LLM-based tools hold considerable promise for the survey design stage, often with comparatively low risk (Rothschild et al. 2024). LLMs, with their ability to quickly interpret, rephrase, and iterate on question design, can act as expert advisors in aligning the wording of survey items with the theoretical objectives of the study, helping to reduce specification error (the failure to capture the underlying theoretical construct). Researchers can prompt an LLM to generate a single survey question within a response domain of interest (Buskirk, Eck, and Timbrook 2024) or a set of questions based on a conceptual brief (Padgett, Maiorino, and Gutierrez 2024; Maiorino et al. 2023), for example: "a 5–10 item battery measuring latent approval of the president," and receive drafts that are structurally sound and varied in framing. Alternatively, researchers can submit existing question sets to the LLM for diagnostic review or pre-testing (Tao et al. 2024), inviting feedback on possible mismatches between survey items and the conceptual model or in generating alternate versions of survey questions that are used repeatedly in longitudinal studies (Yun et al. 2023).

LLMs can also help to reduce measurement error – distortions introduced into responses by how questions are asked or understood. Existing LLM-based tools are capable of evaluating survey questions for clarity, tone, and cultural or linguistic bias. For example, LLMs are now routinely used to assist with language translation. LLMs can be tailored to assess how respondents with different demographic backgrounds or from different countries might interpret a given question (Adhikari et al. 2025). Further, it is increasingly possible to not just have LLMs review just the words of the survey, but have the LLMs explore the survey loaded into the actual infrastructure which the respondents use to answer the questions, to consider possible bias in how the questions that infrastructure may induce beyond the question wording. All of this offers new opportunities for reducing variability correlated with the construct being measured.

The risk of using LLMs in these ways is low. Expert human researchers can always review the models' edits and suggestions. The biggest risk is that LLM-based tooling could give false confidence for non-experts such as over estimates of a survey question's readability (Olson and Buskirk 2025), who may not be able to understand where the model has made a mistake or given them bad advice.

Gains in this area are contingent on further research into the design and tuning of LLMs for these purposes. Key empirical questions remain unanswered: What training data and prompt structures best align LLM outputs with best practices in survey methodology? What data is a given model drawing on for their answers, are the models merely sophisticated parroting of internet content, or are they valid reflections of domain-specific expertise (directly querying ChatGPT, it confirms access to code-books of major public surveys, but does not provide a clear answer to the breadth of survey code-books in its training data[1])? How do LLMs perform across domains (e.g., political attitudes vs. consumer preferences)? Do different uses of LLM-based tools, such as question generation versus critique, introduce distinct behavioral biases or foster over-reliance? And, it is not just the models, but the tooling that will shape the impact: can tools adjust for experts and non-experts (where measurement errors are likely quite different)? Can tools induce the right feedback loop between substantive expertise and survey methodology (plausibly shifting how we understand and address specification error)?

2.2 Administration Stage (AI as the Interviewer): High Potential, High Risk

Even more disruptive is the potential for interactive or conversational surveys, in which LLMs engage respondents in a dynamic dialogue. These novel interfaces have the potential to increase engagement, reduce respondent fatigue, and improve data richness (Barari et al. 2025; Xiao et al. 2020; Wuttke et al. 2025; Velez and P. Liu 2025). However, they also introduce new methodological and

[1] Queried GPT3.5 on May 6, 2025

ethical concerns: does this approach reduce consistency across respondents? What latent cues or biases might the LLM convey through language use, question ordering, or other ways in which the question may influence the response? What kind of consent or transparency is required when using AI-driven conversational agents in place of static survey forms?

The risk of AI involvement here is much higher because LLMs pose questions directly to respondents without human oversight. The model could ask leading or even inappropriate questions, or fail to ask questions on topics the researcher would like to pursue. Additionally, data privacy and security risks arise when sensitive respondent information, plausibly the direct result of questions that probed into sensitive areas or asked for private information, is processed and stored by AI systems. Widespread negative experience by just a few survey firms or researchers (e.g., a few viral examples of AI survey tools being inappropriate going viral) could have meaningful effects across the industry, as it could affect respondent trust and willingness to engage across potential respondent pools.

These developments call for a new research agenda on interactive measurement and open-ended data quality. What safeguards are needed to ensure transparency and replicability in dynamic interviews? What forms of documentation or logging are required to make AI-mediated interactions auditably fair and unbiased? How do respondents perceive AI interviewers compared to human interviewers with respect to trust, response accuracy, and willingness to participate? How does the choice of the visual or aural mode affect responses? Does the increased use of open ended questions increase respondent burden and reduce overall data quality and usefulness? Will AI interviewers struggle to adapt dynamically to unexpected or nuanced responses, potentially leading to misinterpretation or lost opportunities for deeper insights (i.e., should we worry about long-tail conversations that really mess up the interviewers, especially considering potentially malicious actors).

2.3 Administration Stage (AI as the Respondent): Limited Impact, High Risk

One of the more controversial applications of LLMs is the use of synthetic respondents, AI-generated responses designed to simulate human responses. Some view this as a purely cost-saving innovation, while others see potential in reducing coverage or non-response error by capturing the perspectives of underrepresented or missing populations. The appeal is intuitive: if certain segments of a population are inaccessible or unresponsive, perhaps their "voices" could be approximated using large-scale models that have been trained on vast and diverse linguistic data.

Early research offers mixed but intriguing results. Under specific conditions – such as predicting answers to known questions across different demographic combinations, or modeling adjacent questions from the same battery—synthetic respondents, have shown some predictive validity (Brand, Israeli, and Ngwe 2023; Suh et al. 2025). These cases resemble the types of adjustments already handled by multilevel regression with post-stratification (MRP), but with a generative rather than statistical approach (and expanding from unseen demographics to unseen, but related, questions). However, serious limitations remain. Synthetic respondents often fail to replicate the coherence, contextual awareness, and variability found in genuine human responses, particularly across diverse survey formats and topics (Bisbee et al. 2024; Sepulvado, J. Y. Lerner, and Huang 2025; Boelaert et al. 2025; Santurkar et al. 2023; Qu and Wang 2024; Sanders, Ulinich, and Schneier 2023). We know relatively little about where synthetic respondents succeed and where they fail, particularly when it comes to capturing the views of smaller sub-populations, marginalized groups, or topics that evoke distinct cultural, emotional, or experiential dimensions.

Synthetic respondents also inherit the biases of their training data. These biases may be historical (e.g., overrepresentation of majority viewpoints), systemic (e.g., underrepresentation of certain dialects, media, or regions), or behavioral (e.g., mimicking common rhetorical styles). Without strong theoretical grounding and empirical testing, there is a risk that synthetic supplementation may am-

plify misrepresentation rather than correct it, by offering a false sense of coverage while failing to reflect the authentic views of missing populations (introducing ethics concerns, along with the error). We noted above that ChatGPT tells us that it consumes code-books, but it also tells us that it does not consume much individual-level response data from surveys, some aggregated results, but many more write-ups of results.[2] This extraction from the raw response data could create bias in how GPT assumes certain sub-population may respond to surveys.

Further error is potentially introduced in how commercial LLM-models assume the answer to surveys due to how they are fine-tuned to create generally pleasant user experiences. This may lead to more acquiescence bias (the tools want to agree with the questioner), further reduction in variance leading to more modal answers even when their is true variance, and general reduction in non-socially acceptable answers (which may be what the researcher is exploring).

Ironically, a key historical factor reducing the demand for synthetic respondents was the rapid expansion of non-probability online panels in the 2000s and 2010s. As the cost of recruiting and surveying human respondents declined, the economic case for synthetic modeling weakened. In a sense, the last methodological disruption displaced the extreme need for the kind of approach that synthetic respondents now seek to reintroduce, as the cost savings of synthetic respondents is minimized making it less attractive relative to the risk. Yet, many researchers and practitioners may be happy to trade off increased error for less cost. And, paying a model for synthetic responses will be cheaper than human responses, even with the dramatic decrease in costs for human responses over the last few years. Any use of synthetic respondents is the most critical for transparency requirements we discuss below, because it carries the highest risks for end stakeholders.

Even if researchers do not intentionally deploy synthetic responses, there's the risk that human response pools may become contaminated by synthetic data. This could occur if respondents use LLMs to speed through surveys or if automated bots impersonate human respondents. Although this risk falls partly on firms providing respondents, researchers also bear responsibility for implementing best practices to mitigate such risks (Veselovsky et al. 2025; Martherus, Cook, and Podkul 2025).

In short, while synthetic respondents are an intellectually provocative idea, they currently pose high methodological risk, with uncertain practical benefit. Their use also raises ethical concerns, given the potential to mislead stakeholders about the authenticity and representativeness of the data. At present, they should not be viewed as a reliable solution for replicating human responses, even if aimed at correcting coverage or non-response bias. Indeed, given the various issues identified, they could well exacerbate representational bias.

Yet, future research may help mitigate some of these risks. Particularly promising directions include the development of domain-specific fine-tuned models for survey response generation, greater transparency in training data provenance, and systematic benchmarking across survey domains and population strata. Critically, researchers must assess not only aggregate accuracy but also subgroup validity: evaluating whether synthetic responses capture within-group heterogeneity and sociopolitical nuance, rather than merely reflecting the mean of an assumed demographic. And questions are not an in a vacuum, future research is going to need to address with and between-respondent correlations across questions, as these correlations are frequently the target of research.

Beyond the use of synthetic data, we anticipate that AI will, in the longer-term, increasingly enhance the efficiency of survey targeting and response, potentially reducing both coverage and non-response error. Targeting procedures may be developed around autonomous agents or other digital tools capable of dynamically identifying and engaging relevant respondent pools – representing a significant and ongoing disruption to traditional survey methodologies. Additionally, real-time data processing

[2] Queried GPT3.5 on May 6, 2025

and respondent targeting could facilitate iterative sampling strategies. One promising yet speculative avenue involves the integration of LLMs into adaptive survey systems that adjust sampling decisions or survey logic in response to interim data quality assessments. For instance, LLMs may be employed to detect emerging inconsistencies during data collection, enabling responsive modifications to sampling strategies or question ordering to mitigate total survey error. Moreover, recent evidence suggests that LLM-generated messaging can be effective in persuasive contexts (Salvi et al. 2024), raising the possibility that such models could optimize respondent engagement and improve response rates. Nonetheless, these applications remain nascent and pose considerable methodological, ethical, and operational challenges. Realizing their potential will require the development of robust theoretical frameworks, rigorous empirical validation, and the creation of infrastructure to support real-time monitoring and intervention.

2.4 Analytics Stage (AI as a Labeler): High Potential, Medium Risk

LLMs enable scalable analysis of open-ended responses, including real-time summarization and categorization that can be integrated into data collection (Allamong, Jeong, and Kellstedt 2025; Karousos et al. 2024; Mellon et al. 2024). This ability lowers the cost of including open-text items, frequently avoided due to cost, coding complexity, and interpretive ambiguity. Respondents could even choose to read or hear the questions, and respond by typing or speaking, as LLM transcription lowers the cost of mixed-mode administration of surveys. These advances have the potentially to radically update the design of surveys.

There is meaningful risk in AI-analysis of open-ended questions, both on the stability of the findings (forms of specification, measurement, and processing error) and ethical implications (Bedemariam et al. 2025). First, simple variations of prompts and models can produce very different individual-level and aggregated labeling for a given set of open-ended answers (i.e., there is a lot of potential processing error in the instability of the derived answers). This is also sensitive to the exact goal of the researcher from the open-ended answers (opening up questions about both specification and measurement error). For example a researcher could get a group of open-text responses about politics and derive: the political lean of the respondents, top narratives, if certain narratives (derived elsewhere) are present, how much a topic is mentioned, the sentiment of the answers, or the sentiment towards the main agent in the answers. The scale and flexibility of the open answers is exciting, but using the same answers to explore multiple very different questions carries a lot of risk of error. Second, there are meaningful ethical concerns with plugging open-ended survey responses into commercial LLM-based tools, where there is a lot of unknown variation in how the data could be stored or used in the future.

These developments call for a new research agenda on interactive measurement and open-ended data quality. For instance: How does LLM summarization of open responses vary by prompt, topic, population, or model version? How can the research community evaluate the quality of labels generated by LLMs in the absence of a definitive ground truth: Is it sufficient to have human reviewers assess the labels by comparing them to a sample of the open-text responses, or should we require that humans generate a small set of labels to compare with those derived by the LLM?

2.5 Analytics Stage (AI as the Modeler): High Potential, Medium Risk

Among the various components of total survey error, processing error – those errors introduced after data collection during the cleaning, coding, transformation, and analysis of survey data – may be the area where LLMs currently offer the most immediate and tangible benefit (but with associated higher risk depending on the situation). These errors can stem from inconsistent coding, undocumented data transformations, manual entry mistakes, or flawed logic in post-field processing. Unlike coverage or non-response error, which require structural interventions at the design or sampling stage,

processing errors are largely procedural, making them well-suited to automation.

LLMs can aid in a range of processing tasks. They can assist with variable recoding, flag logic inconsistencies, generate metadata (e.g., variable descriptions or survey flow documentation), and summarize patterns in open-ended responses (as discussed above) (Allamong, Jeong, and Kellstedt 2025; Olivos and M. Liu 2024; J. Lerner et al. 2024; Wong 2024). They also offer an underexplored opportunity to simulate pilot datasets, which researchers can use to stress-test coding scripts, check statistical models, or validate survey logic prior to full deployment (Rothschild et al. 2024). These applications have the potential to improve both efficiency and reproducibility, particularly in contexts where complex survey instruments and multi-stage workflows can introduce subtle but consequential errors.

However, these gains come with risks. Automation of data handling using LLMs introduces a new layer of opacity into the research. If researchers rely on model-generated transformations without thorough documentation or validation, they may inadvertently obscure the provenance of their results. For example, if a model reclassifies open-ended responses into categorical codes, but the criteria for that classification are unclear or unreproducible, downstream analyses could be distorted in ways that are difficult to detect or correct. Worse, errors introduced by LLMs may appear authoritative, given the fluency and confidence of their outputs, leading to false confidence in flawed results.

As with other areas of survey research, the solution is not to avoid LLMs, but to develop guardrails and best practices around their use. This includes maintaining clear audit trails of prompts and outputs, establishing criteria for human validation, and incorporating quality checks (e.g., intercoder reliability metrics for model-generated codes). Importantly, LLMs should be integrated into human-in-the-loop systems, where researchers have ultimate responsibility and oversight for data processing decisions.

In sum, LLMs present significant opportunities to reduce processing error through automation, consistency, and scalability. Yet these tools must be deployed with caution. Without transparency and validation, LLM-driven processing may become a new source of error rather than a remedy. Future research should explore not just where these tools can assist, but how they can be integrated responsibly into existing workflows, and under what conditions they meaningfully improve data quality without compromising transparency or replicability.

2.5.1 Reporting Stage (AI as the Briefer): High Potential, Manageable Risk

Survey results are frequently disseminated through static formats such as presentations or reports, which represent a curated distillation of the information provided by respondents. While these formats are useful for summarization, they often hinder the ability to perform longitudinal analyses or draw systematic comparisons across different survey iterations. However, recent advancements in large language model (LLM)-based tools present an opportunity to build systems that enable natural language querying and comparison of survey data.

For instance, chatbots can enable analysis of survey datasets using natural language, converting raw survey data into narrative reports that present results in plain language or data visualizations. Automated translation could serve to expand access to global data sources. AI tools could be designed to more easily integrate external datasets to support comparative analysis and enhance the validation of results. Such systems open up possibilities for more accessible, flexible, and dynamic engagement with data with a wider range of data users.

While there are a handful of proprietary tools, several research challenges remain for widespread adoption. It is crucial to ensure that such systems do not mislead end users by surfacing spurious correlations – issues that trained researchers typically account for through rigorous modeling and statistical testing. Addressing this also involves improving data literacy among non-expert users, helping them to engage critically with statistical results.

3 Survey Research Field Taking Ownership of their Disruptions

The integration of LLMs into survey research represents a methodological turning point, one that may well surpass the disruptive transition from telephone-based probability sampling to online non-probability methods in the early 2000s. That earlier disruption was driven by the search for cost efficiency and operational flexibility as traditional telephone surveys became prohibitively expensive and less effective. The shift brought with it not only new modes of data collection but also experimentation in survey question design (as instruments adapted to online platforms) and new analytic techniques (to address increasingly unbalanced samples).

The rise of non-probability surveys in both academic and practical surveys was not handled as well by the survey research field. The concerns raised by survey researchers during that time were real and pressing, especially regarding sample quality, non-coverage, and bias. But the response was hindered by inconsistent disclosure practices, weak norms, and a general lack of shared methodological frameworks (Baker et al. 2010; Callegaro et al. 2013). Journals often struggled to evaluate research based on non-probability samples, and reviewers – sometimes justifiably, sometimes not – met such studies with skepticism or outright hostility. Rather than adapting and integrating these emerging data sources, some within the field rejected them outright, creating a rift within research and practice.

This fragmentation that occurred within the field over non-probability samples had long-lasting effects. As a result of impaired access, limited understanding, and institutional inertia, the research community initially failed to build the standards and infrastructure needed to manage the shift. This lack of early cohesion led to several significant and enduring harms. Trust in survey research eroded as researchers and practitioners struggled to agree on best practices, contributing to skepticism among survey consumers. It became increasingly difficult to know which surveys could be trusted and for what purposes, further complicating the landscape for both researchers and practitioners. The absence of a unified approach also hindered methodological innovation, as researchers were often working in silos, unable to build on each other's work effectively, which slowed the overall progress in the field. It is our concern, as researchers in the field, that fragmentation and resulting inconsistencies in survey research practices led to a diminished perceived value of survey research, with stakeholders questioning the validity and reliability of survey data, impacting its influence on policy and decision-making.

We now face a similar inflection point – this time driven not by changes in sampling frames, but by the introduction of artificial intelligence. To avoid repeating past mistakes, the adoption of LLMs must be accompanied by a proactive, field-wide effort to develop governance frameworks, norms of disclosure, and shared infrastructures for evaluation. A smooth transition into this AI-mediated era requires two core commitments:

Transparency: Researchers must clearly disclose when and how LLMs are used throughout the research lifecycle – from question design to data collection, analytics, and reporting.

Standards: The field must collectively build adaptive, flexible standards that can evolve alongside AI capabilities, ensuring that methodological rigor, transparency, and ethical responsibility scale with innovation.

To operationalize these goals, we propose the creation of a commission or task force, ideally under the auspices of AAPOR, to develop best practices for the transparent and ethical use of LLMs in survey research. This group could extend AAPOR's existing Transparency Initiative (American Association for Public Opinion Research n.d.), incorporating specific guidelines for the use of generative AI and LLMs in survey pipelines. Its output would serve academics, industry professionals, and policymakers alike: anyone who relies on credible survey data for empirical insight or decision-making.

The goal here is to ensure that any researcher reading the survey results could understand the process enough to replicate it (accepting that just like human coding, some aspects of the replication process would not necessarily get the same answers due to the randomness and evolution of the coders or models and the passage of time), and that surveys conform to our current understanding of basic quality and ethical guidelines (which need to evolve with the research and technology).

As a first step, we recommend that all researchers and organizations reporting survey findings involving LLMs include documentation along the following dimensions.

1. **Use of LLMs:** Was an LLM used to generate, revise, or evaluate survey items? Did it interact directly with respondents (e.g., via chatbots or conversational survey modes)? Did it create any synthetic responses? Was the model used to summarize open-ended responses, assist with coding, or interpret patterns in the data? Authors should specify the model used (e.g., GPT-4, Claude, LLaMA), its version or checkpoint, the date of access, and the scope of application. This will allow researchers to understand the potential risks and limitations.

2. **Prompt transparency and human oversight:** Where applicable, authors should share sample prompts or prompt templates, description of the prompting approach (e.g. zero shot, few-shot, chain of thought, etc.) along with a description of human oversight – e.g., whether outputs were manually reviewed, edited, or validated, and by whom. This helps surface interpretive choices and highlight where expert judgment shaped the results. As our field continues to advance the integration and use of LLMs, sharing prompts that didn't work and why may be as important as sharing prompts that did work.

3. **Validation:** Where applicable, LLM inference should be validated. In some use cases, such as qualitative coding, human raters can score AI responses on accuracy and relevance. Others require more systematic, statistical comparisons to benchmarks, not just distributions, but also for subgroups. Understanding edge cases also requires adversarial testing – intentionally providing challenging, ambiguous, or misleading inputs with the goal of uncovering biases that might not appear under normal testing conditions.

To build momentum and establish accountability, we urge journals such as *Public Opinion Quarterly*, *Survey Research Methods*, *Survey Methodology*, and *Journal of Survey Statistics and Methodology* to incorporate AI-specific transparency standards into their author guidelines. Establishing a baseline expectation for disclosure now will accelerate convergence on shared norms, providing valuable data on what researchers are doing with AI to help spearhead standards, and signal to both researchers and practitioners that the field is serious about methodological integrity in the age of AI. These standards should be living documents, revised over time through empirical testing, community input, and model evaluations.

Crucially, this governance effort should not be viewed as a brake on innovation, but as its enabler. Responsible disclosure and shared standards are essential for building confidence in new methods, scaling their legitimacy, and sustaining public trust. LLMs offer immense promise–from reducing measurement error to enabling richer, more adaptive survey designs–but this promise will only be realized if the research community takes ownership of its own disruptions. We have been here before. We know what happens when we resist change, and we know what is possible when we face it head-on. The time to build the infrastructure for the LLM era is now.

3.1 Interdisciplinary cooperation

Behavioral science has already seen a lot of work on the impact on LLMs, not just the immediate augmentation of current research practices, but more fundamental disruptions to the work (Horton 2023; Manning, Zhu, and Horton 2024). The survey research and behavioral science communities

share a lot of infrastructure, from audience companies, to survey tools, to analytical methods. And, there are several other fields in a similar position, using survey tools for related work. Cooperation could be mutually beneficial as we all face similar challenges.

4 Survey Research Field Working with AI Community

It is one thing for the survey research community to come together and develop internal norms, standards, and governance, but to fully realize the potential of LLMs while minimizing their risks, this effort must be extended outward. The next phase of this transition will require new partnerships between the survey research community and the AI community, including computer scientists, NLP researchers, tool developers, and platform designers. These are not just methodological conversations, they are infrastructural collaborations.

As discussed in the previous sections, LLMs show high potential in automating parts of the data processing pipeline, reducing human error, and enabling more scalable and reproducible workflows. But for these gains to be realized responsibly, the tools themselves must evolve in ways that reflect the needs and values of survey research. This includes features like transparent logging, reproducible outputs, support for human-in-the-loop workflows, model versioning, domain fine-tuning, and subgroup evaluation capabilities. Without direct engagement, LLM tooling will continue to be shaped primarily by adjacent domains, such as marketing, customer service, or product feedback analysis, where stakes, standards, and workflows are very different.

The history of methodological disruption in survey research, particularly the uneasy transition from probability to non-probability sampling, shows the danger of falling behind the tools that shape the practice. When researchers are not at the table during the design of new platforms or technologies, they are often forced to play catch-up, reacting to changes they did not anticipate and trying to retrofit accountability onto systems never designed for it. That reactive posture, as we have seen, can delay the development of best practices, fracture consensus, and undermine trust.

Avoiding that fate in the LLM era means building active, intentional collaborations with the AI community:

1. **Survey methodologists and social scientists** can help AI researchers understand the unique constraints of survey design, including the importance of question wording, framing effects, representativeness, and interpretability (Eckman, Plank, and Kreuter 2024).

2. **AI researchers and developers** can help embed technical affordances, such as interpretability tools, documentation protocols, and auditability, into LLMs and associated survey platforms. These researchers may also help survey researchers broaden the presentation of public use data so that it is more machine readable for consumption in future training of LLM models.

3. **Shared research agendas** can help ensure that benchmarks and evaluation methods reflect the actual use cases and error models relevant to survey work (e.g., specification and measurement error), not just general-purpose language tasks.

There are promising precedents here. Recent collaborations between social scientists and machine learning researchers have produced fine-tuned models for marketing and political estimation (Brand, Israeli, and Ngwe 2023; Suh et al. 2025), and new methods for classifying open-ended responses using supervised and unsupervised NLP. But these are still isolated efforts. What is needed now is a more durable interface between communities: dedicated workshops, cross-disciplinary fellowships, co-authored benchmarks, and shared tooling initiatives. The BigSurv (www.bigsurv.org) conference series is one example of this type of interface with conferences in 2018, 2021 and 2023 bringing survey and social scientists together with data and computer scientists.

Just as we've proposed governance and transparency standards within the survey field, we now call for institutional bridges to the AI world. This could include joint initiatives between AAPOR and conferences like ACL, EMNLP, or NeurIPS, or coordinated funding opportunities from agencies like NSF and the SSRC to support interdisciplinary teams.

Ultimately, the design of survey workflows and the design of AI tools should not proceed on parallel tracks. They are now part of the same ecosystem. The tools that shape how questions are written, how responses are analyzed, and how data are processed must be responsive to the principles of valid, representative, and interpretable survey research. And those principles must now be legible – and actionable –to the engineers and scientists building the next generation of AI systems.

Bringing these communities into dialogue is not just an aspirational goal – it's a practical necessity for ensuring that LLMs become a force for rigor, rather than risk, in the future of survey research.

5 Discussion

Our call to action is for the community to closely **monitor and evaluate LLM use cases** across the survey pipeline to understand their evolving impact on data quality and ethical considerations. Key to this effort is to identify research gaps and prioritize funding for innovations. This will help us build metrics that **evaluate models and tools**, and **establish evaluation frameworks and benchmarks** to guide progress. With that understanding we advocate to incentivize **transparency and disclosure**, including addressing conflicts of interest, and **adapt best practices** to keep pace with technological developments. We aim to **foster both cross-discipline and industry-academic partnerships** that bring together diverse expertise and include adjacent stakeholders, such as marketers, journalists, and social scientists, to ensure a holistic view of the implications of LLM integration and potential solutions on both the research and tooling. Additionally, we advocate for training both researchers and consumers, ensuring they have the knowledge and tools to navigate this dynamic landscape responsibly.

The integration of LLMs into survey research presents both transformative opportunities and significant challenges. As outlined in this paper, AI tools, particularly LLMs, have the potential to improve nearly every phase of the survey lifecycle, from question design and administration to data processing and analysis. However, these advancements come with considerable risks, including the potential amplification of bias, reduced transparency, and a risk of fragmentation within the research community that could undermine the quality and credibility of survey research.

One of the central findings of our analysis is the uneven impact of AI across different areas of survey methodology. Certain aspects of survey research, such as the mitigation of specification and measurement errors with design and analytical uses of LLMs, appear to offer substantial benefits with comparatively manageable risks, especially when robust human oversight is in place. In contrast, applications like synthetic respondents or dynamic, conversational surveys involve higher methodological uncertainty and pose more serious ethical concerns. These differences highlight the importance of adopting a nuanced, domain-specific approach to the integration of AI in survey work.

A useful parallel can be drawn to the advent of non-probability sampling. Although that methodo logical shift ultimately spurred innovation, it was initially characterized by limited coordination, inconsistent standards, and a decline in public trust. The field now faces a similar inflection point. The choices made in the coming years – by researchers, institutions, journals, and technology developers – will not only shape the trajectory of AI in survey research but also influence the legitimacy and societal value of the field as a whole.

Future research should focus on empirically validating AI applications across a range of survey contexts, paying particular attention to subgroup validity, reproducibility, and ethical safeguards. Equally

important is the development of both technical and institutional infrastructure to support responsible experimentation and accelerate collective learning.

The integration of AI into survey research is not a matter of if, but how. The measures we advocate should be seen not as constraints on innovation, but as foundations for responsible, sustainable progress. By approaching this transition with clarity, humility, and a commitment to shared responsibility, the research community can ensure that AI becomes a driver of rigor, inclusivity, and innovation, rather than a source of risk and fragmentation.

References

Adhikari, D. M. et al. (2025). "Exploring LLMs for Automated Pre-Testing of Cross-Cultural Surveys". In: *arXiv preprint arXiv:2501.05985*.

Allamong, M. B., J. Jeong, and P. M. Kellstedt (2025). "Spelling correction with large language models to reduce measurement error in open-ended survey responses". In: *Research & Politics* 12.1.

American Association for Public Opinion Research (n.d.). *AAPOR Transparency Initiative*. https://www.aapor.org/Standards-Ethics/Transparency-Initiative.aspx.

Argyle, L. P. et al. (2023). "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* 31.3, pp. 337–351. DOI: 10.1017/pan.2023.2.

Baker, R. et al. (2010). "Research Synthesis: AAPOR Report on Online Panels". In: *Public Opinion Quarterly* 74.4, pp. 711–781. DOI: 10.1093/poq/nfq048.

Barari, S. et al. (2025). *AI-Assisted Conversational Interviewing: Effects on Data Quality and User Experience*. arXiv: 2504.13908 [cs.HC]. URL: https://arxiv.org/abs/2504.13908.

Barrie, C., A. Palmer, and A. Spirling (2024). "Replication for Language Models Problems, Principles, and Best Practice for Political Science". In: *URL: https://arthurspirling.org*.

Bedemariam, R. et al. (2025). *Potential and Perils of Large Language Models as Judges of Unstructured Textual Data*. URL: https://arxiv.org/abs/2501.08167.

Bisbee, J. et al. (2024). "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models". In: *Political Analysis* 32.4, pp. 401–416. DOI: 10.1017/pan.2024.5.

Boelaert, J. et al. (2025). "Machine Bias. How Do Generative Language Models Answer Opinion Polls?" In: *Sociological Methods & Research*, p. 00491241251330582.

Brand, J., A. Israeli, and D. Ngwe (2023). "Using LLMs for market research". In: *Harvard Business School Marketing Unit Working Paper* 23-062.

Buskirk, T. D., A. Eck, and J. Timbrook (2024). "The Task Is to Improve the Ask: An Experimental Approach to Developing Optimal Prompts for Generating Survey Questions from Generative AI Tools." In: *79th Annual AAPOR Conference*. AAPOR.

Callegaro, M. et al. (2013). *Online Panel Research: A Data Quality Perspective*. Chichester, UK: Wiley. ISBN: 9781119941763.

Eckman, S., B. Plank, and F. Kreuter (2024). "Position: insights from survey methodology can improve training data". In: *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. Vienna, Austria: JMLR.org.

Eloundou, T. et al. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*. arXiv: 2303.10130 [econ.GN].

Groves, R. M. and L. Lyberg (2010). "Total survey error: Past, present, and future". In: *Public opinion quarterly* 74.5, pp. 849–879.

Heyde, L. von der, A.-C. Haensch, and A. Wenz (2025). "Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice". In: *Social Science Computer Review*.

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.

Karousos, N. et al. (2024). “A Hybrid Text Summarization Technique of Student Open-Ended Responses to Online Educational Surveys”. In: *Electronics*. URL: https://api.semanticscholar.org/CorpusID:272784277.

Lerner, J. et al. (2024). “Assessing the Accuracy of and Bias with Zero-Shot Text Classification using GPT: A Case Study with Social Media and Survey Data”. In: *Federal Computer Assisted Survey Information Collection Workshops*.

Lyberg, L. (2012). “Survey quality”. In: *Survey Methodology* 38.2, pp. 107–130.

Maiorino, A. et al. (2023). “Application and evaluation of large language models for the generation of survey questions”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 5244–5245.

Manning, B. S., K. Zhu, and J. J. Horton (2024). *Automated social science: Language models as scientist and subjects*. Tech. rep. National Bureau of Economic Research.

Martherus, J., E. Cook, and A. Podkul (2025). “Are Bots Taking Online Surveys?” In: *80th Annual AAPOR Conference*.

Masterton, M., A. Henriques, and D. Mclean (2025). “Why Synthetic Research Is the Next Phase of Human Understanding and Decision Making”. In: *American Association for Public Opinion Research (AAPOR)*. St. Louis, MO.

Mellon, J. et al. (2024). “Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale”. In: *Research & Politics* 11.1.

Olivos, F. and M. Liu (2024). “ChatGPTest: Opportunities and Cautionary Tales of Utilizing AI for Questionnaire Pretesting”. In: *Field Methods* 0.0.

Olson, K. and T. D. Buskirk (2025). “Chatbot Is a Two-Syllable Word...or Is It? Using Generative AI for Survey Question Readability Assessments”. In: *American Association for Public Opinion Research (AAPOR)*. St. Louis, MO.

Padgett, Z., A. Maiorino, and S. Gutierrez (2024). “Evaluating the Quality of Questionnaires Created with Surveymonkey’s Build with AI”. In: *79th Annual AAPOR Conference*. AAPOR.

Qu, Y. and J. Wang (2024). “Performance and biases of Large Language Models in public opinion simulation”. In: *Humanities and Social Sciences Communications* 11.1, pp. 1–13.

Rothschild, D. M. et al. (2024). “Opportunities and risks of LLMs in survey research”. In: *Available at SSRN*.

Salvi, F. et al. (2024). “On the conversational persuasiveness of large language models: A randomized controlled trial”. In:

Sanders, N. E., A. Ulinich, and B. Schneier (2023). “Demonstrations of the potential of AI-based political issue polling”. In: *arXiv:2307.04781*.

Santurkar, S. et al. (2023). “Whose opinions do language models reflect?” In: *International Conference on Machine Learning*. PMLR, pp. 29971–30004.

Sepulvado, B., J. Y. Lerner, and L. Huang (2025). “LLMs Do Not Respond like Humans: Experiments in Model Fine Tuning”. In: *American Association of Public Opinion Research annual meeting*.

Suh, J. et al. (2025). “Language model fine-tuning on scaled survey data for predicting distributions of public opinions”. In: *arXiv preprint arXiv:2502.16761*.

Tao, R. et al. (2024). “Using Large Language Models (LLM) to Pretest Survey Questions”. In: *American Association for Public Opinion Research (AAPOR)*. Atlanta, GA.

Velez, Y. R. and P. Liu (2025). “Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments”. In: *American Political Science Review* 119.2, pp. 1036–1053.

Veselovsky, V. et al. (2025). “Prevalence and prevention of large language model use in crowd work”. In: *Communications of the ACM* 68.3, pp. 42–47.

Wong, W. (2024). “Using Large Language Models for Other-Specify Coding”. In: *Federal Computer Assisted Survey Information Collection Workshops*.

Wuttke, A. et al. (2025). *AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers*. arXiv: 2410.01824 [cs.HC]. URL: https://arxiv.org/abs/2410.01824.

Xiao, Z. et al. (June 2020). "Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions". In: *ACM Trans. Comput.-Hum. Interact.* 27.3. ISSN: 1073-0516. DOI: 10.1145/3381804. URL: https://doi.org/10.1145/3381804.

Yun, H. S. et al. (2023). "Keeping Users Engaged During Repeated Administration of the Same Questionnaire: Using Large Language Models to Reliably Diversify Questions". In: *arXiv preprint arXiv:2311.12707*.