Responsible AI Integration in Survey Research

David M. Rothschild (Microsoft Research)*, Jenny Marlar (Gallup)*, Ashley Amaya (Pew), Soubhik Barari (NORC at the University of Chicago), Trent Buskirk (Old Dominion University), Curtiss Cobb (Meta), Jen Gennai (T3), Sunshine Hillygus (Duke), Ramya Korlakai Vinayak (University of Wisconsin-Madison), Masha Krupenkin (University of Maryland), Sunghee Lee (University of Michigan), Darby Steiger (SSRS), Brock Webb (US Census Bureau)

*Co-Chairs

The Task Force would like to thank Rene Bautista, Stephen Blumberg, Mario Callegaro, Stephanie Eckman, Emily Geisen, Krista Jenkins, Michael Link, Brenna Muldivan, Stanley Presser and Henning Silber for their invaluable review and feedback. The Task Force would also like to thank Hannah Cha for her support of the report and development of the disclosure tool.

Suggested citation: Rothschild, D., Marlar, J., Amaya, A., Barari, S., Buskirk, T., Cobb, C., Gennai, J., Hillygus, D., Korlakai Vinayak, R., Krupenkin, M., Lee, S., Steiger, D., Webb, B.(2026). *Responsible AI Integration in Survey Research.* American Association for Public Opinion Research (https://aapor.org/wp-content/uploads/2026/05/Responsible-AI-Integration-In-Survey-Research.pdf).

Artificial Intelligence (AI), particularly large language models (LLM) and other generative systems, is rapidly transforming survey research across the full research life cycle: from question design and data collection to analysis and dissemination. While algorithmic tools have long influenced survey methods, recent advances in scale, flexibility, and accessibility raise novel methodological, ethical, and governance challenges that existing standards do not fully address. This report, produced by the AAPOR Task Force on Responsible AI Integration in Survey Research, provides a structured framework for understanding where and how AI is being used in surveys, evaluates associated benefits and risks, along with necessary research and development, through the lens of total survey error (TSE) and human–subject protections, and offers guidance for responsible practice. We propose principles for evaluation and human oversight and use that to introduce a practical and tractable disclosure framework designed to promote transparency, informed interpretation, and reproducibility of AI-enabled survey research. Together these contributions aim to support innovation while safeguarding data quality, trust, and credibility of survey-based research.

Table of Contents

1. Introduction and Executive Summary

Artificial intelligence (AI), especially large language models (LLMs) and other generative systems, is rapidly reshaping how survey research[1] is designed, conducted, analyzed, and communicated. While algorithmic tools have long been part of the survey research ecosystem, recent advances in accessibility, generality, and fluency represent a shift that warrants new methodological, ethical, and transparency standards. This report provides a framework for understanding where AI is already being used in survey research, where its use is emerging, and how the field can integrate these tools responsibly while preserving methodological integrity, participant trust, and public confidence in survey-based evidence.

This report is based on four key premises (Built on Section 2's Brief Background of AI):

1. **AI is a general-purpose research technology, not a targeted survey-specific tool.**
   AI is not designed specifically for survey research; it is a general-purpose technology. As a result, its risks and benefits depend less on whether AI is used at all, and more on how, where, and with what level of oversight it is deployed across the survey lifecycle.
2. **With current tools and technology, usage is more directed at augmentation than full automating.**
   AI is primarily assisting researchers by speeding ideation, dynamically evolving how questions are asked, and powering analytics, rather than replacing core human judgment. The most disruptive effects are likely still ahead.
3. **This report is not the end of the discussion, but the beginning.**
   The risks, benefits, and appropriate disclosures associated with AI-enabled survey research will necessarily evolve as the technology matures. Accordingly, this report is best read as a framework for interpreting current practices and anticipating how methodological standards may need to adapt over time.
4. **Traditional survey quality principles remain valid, but we need to adapt them for our dynamic and rapidly evolving technological environment.**
   Concepts such as total survey error, validation, inter-rater reliability, and disclosure still apply, but must be adapted to accommodate non-deterministic models, opaque systems, and unpredictable model updates.

Section 3: AI's Use in Surveys

The report categorizes AI's emerging roles across the survey workflow by functional risk, rather than by the traditional linear sequence of design, collection, analyzing, and briefing. Ordering these roles from higher to lower risk (collection to analyst to briefer to colleague) helps clarify where more research, heightened transparency, and increased oversight are most essential. Risk varies substantially within each role depending on the scope of the task, the degree of autonomy afforded to AI, and the extent to which human review remains feasible.

[1] For the purposes of this report, the term "survey research" is used comprehensively to encompass both quantitative and qualitative research.

AI as a Data Collector (Data Collection): Although still relatively uncommon, AI is increasingly explored as a means of administering surveys, either through conversational interviewers or in the form of synthetic responses. Conversational agents may improve engagement or respondent comprehension in some contexts, but they complicate standardization and raise new concerns about auditability, neutrality, consent, and measurement error. Synthetic responses are used to supplement or replace human respondents and pose particularly serious validity and disclosure risks if applied beyond clearly labeled pretesting, pilot work, or exploratory diagnostics.

AI as an Analyst (Post–Data Collection Processing and Modeling): AI is now widely used for transcription, translation, data cleaning, coding of open-ended responses, thematic clustering, and exploratory modeling. In some labeling and classification tasks LLMs can approach or match human performance while substantially reducing cost and latency. At the same time, instability across prompts and model versions, domain-specific biases, and opaque transformations introduce novel sources of processing error. These risks make explicit documentation, sensitivity analysis, and human validation essential.

AI as a Briefer (Reporting and Dissemination): AI systems are beginning to transform how survey findings are summarized and consumed, shifting from static reports toward interactive, natural-language interfaces. These tools can improve accessibility and speed of insight, particularly for non-technical audiences. However, they also increase the risk of oversimplification, spurious inference, hallucinations, and misuse if summaries abstract away uncertainty, caveats, or methodological context. Careful constraint, framing, and audience-appropriate design are therefore critical.

AI as a Colleague (Pre-Data Collection Design and Ideation): AI is increasingly used to draft, edit, test, and translate questionnaires, protocols, and related research materials. These applications offer high potential productivity gains with comparatively lower risk, as outputs are typically reviewed and revised by human experts prior to fielding. Nonetheless, risks remain, including over-delegation of researcher judgment, subtle biases inherited from training data, and unwarranted confidence in AI-generated question wording or structure.

AI as a Workflow (End-to-End Systems): Almost by definition, overlapping AI systems in multiple roles substantially increases risks, but we list this last, as it is still extremely rare in practice. AI systems may increasingly operate across multiple stages of the survey lifecycle, enabling adaptive, iterative, and potentially AI-first workflows. Such systems could meaningfully reduce cost and cycle time, but they also introduce substantial risks to reproducibility, comparability, traceability, and governance. Without robust human-in-the-loop checkpoints, versioning, and explicit transparency mechanisms built in from the outset, end-to-end automation risks obscuring decision pathways and undermining confidence in resulting inferences.

Section 4: Evaluating the Usage of AI for Research

We introduce a practical framework for evaluating data quality and sources of error when using generative AI systems, including LLMs, in survey and social science research. The framework emphasizes that LLMs must be assessed in relation to the specific task they are intended to perform, as performance in superficially similar applications does not ensure validity in new contexts. It organizes evaluation around four core criteria: validity, performance, sensitivity, and reliability. Together, these criteria address whether the model targets the intended task, performs it effectively, holds up under variation in prompts, inputs, or models, and produces substantively consistent results under repeated conditions. Conceptually aligned with established approaches such as total survey error and fit-for-purpose evaluation, the framework does not prescribe universal benchmarks, but instead provides a structured set of considerations to guide task-specific, context-dependent assessment of AI-assisted research.

Section 5: Recommendations for Transparency and Reporting

A central deliverable of this work is an update to disclosure standards for the use of AI in surveys, designed to support transparency, reproducibility, and understanding. The framework specifies:

- **Required disclosures:** the tasks for which AI was used, the number of human respondents, and validation and human oversight. These disclosures represent the minimum information necessary for interpretability, bias assessment, and informed evaluation of results.
- **Enhanced disclosures:** additional documentation intended for studies that require reproducibility, external validation, or deeper auditability of AI-assisted processes.

The checklist is intentionally designed to be practical, focusing on disclosures that meaningfully support understanding and risk assessment; extensible, allowing adaptation as technologies evolve; and interoperable with existing survey reporting standards. Its aim is to integrate smoothly into established disclosure norms to avoid creating a duplicative reporting burden.

Section 6: Responsibility to Human Subjects

The report reinforces that AI does not diminish researchers' ethical obligations to human subjects. Instead, it heightens them. Disclosure, consent, data protection, risk assessment, and fairness must explicitly account for AI-enabled inference, automation, and downstream reuse. Borrowing from established human-subject ethics frameworks, the task force argues for proportional safeguards, continuous reassessment, and participant-centered design as AI capabilities evolve.

Note to Readers

This report is written for survey creators. While it necessarily engages with the responsibilities and practices of vendors, it is not directed at vendors themselves. Nor is it intended for a general-public audience. Although readers outside this core group may find relevant takeaways, the report is deliberately scoped to a specific set of decision-makers and practitioners. Attempting to address all possible audiences would dilute both clarity and impact; accordingly, we remain focused on the needs and constraints of those who design, field, and interpret survey data.

We also recognize the wide range of backgrounds, contexts, and levels of expertise among survey creators. Some readers will bring deep technical knowledge of AI systems, while others may be encountering these tools for the first time through this report. Surveys are also developed for many different contexts, each with distinct goals and constraints. These include internal organizational use, client-facing research, large-scale public deployments, and academic inquiry. To accommodate this diversity, the report is structured to support selective reading. Readers are encouraged to move directly to the sections and subsections most relevant to their work. If using the interactive version, this is aided by an accordion-style framework that clearly delineates topics. We also highlight where expectations around documentation, disclosure, and methodological detail may appropriately differ across publication types and use cases.

This report also makes use of a number of technical and AI-industry-specific terms that may be unfamiliar to some readers. A glossary defining these terms is provided at the end of the document. An interactive version of this report is also forthcoming, in which glossary terms will be hyperlinked throughout the text for easier reference.

We note that this report cites a substantial number of preprints. In many cases, these papers have since been published in peer-reviewed computer science conferences, but the preprint remains the most accessible and widely cited version. More broadly, the pace of technical development in AI far outstrips the traditional academic publication cycle, making preprints an unavoidable component of any timely assessment of the state of AI usage. Wherever possible, we sought to ensure that cited preprints were well cited and influential. Their inclusion reflects not a lack of rigor, but the practical realities of documenting a rapidly evolving field.

Overall Conclusion

AI will not replace survey research, but it will profoundly reshape it. Used responsibly, AI can improve efficiency, expand methodological reach, and lower barriers to insight. Used carelessly or opaquely, it risks undermining validity, trust, and the credibility of survey-based evidence. This report positions AAPOR and the broader research community to navigate that tension by articulating principles, frameworks, and practical tools that allow innovation to proceed without sacrificing the core values of the field.

2. Brief Background on AI

Artificial intelligence (AI) has become increasingly visible across research and professional workflows, including survey research. While AI has long been used in adjacent domains such as automated coding, record linkage, and statistical learning. Recent acceleration of capabilities, accessibility, and scale has substantially expanded its scope of usages (see Appendix A1.2). In particular, advances in large language models (LLMs), a prominent class of generative AI (GenAI), have made it possible for non-technical users to draft, summarize, classify, and transform text through natural-language interaction (Appendix A1.2.1). This section provides a concise orientation to contemporary AI, establishing a shared frame for understanding what these systems are, how they are typically deployed, and why they warrant careful attention in survey research.

A central organizing idea of this report is that AI systems are best understood as layered sociotechnical systems rather than monolithic tools. Their behavior depends on interacting components that span the model layer, training and fine-tuning regimes, prompting strategies, deployment infrastructure, application or agent layers, and governance arrangements (Appendix Background on AI (Full); A1.2). These design choices shape error patterns, bias, reproducibility, transparency, and oversight. In practice, many research teams interact with AI through third-party platforms or APIs, which can further complicate evaluation by obscuring details about model updates, data sources, or operational constraints. As a result, assessing AI systems requires attention not only to outputs, but to how systems are configured and embedded in workflows.

Another key distinction concerns how AI is used. Evidence to date suggests that the strongest and most reliable benefits arise when AI systems augment human judgment rather than automate tasks end-to-end (Appendix A1.2.3). In survey research, augmentation includes assisting with questionnaire drafting, summarizing literature, coding open-ended responses, synthesizing qualitative data, and supporting exploratory analysis: applications that align with AI's strengths in pattern recognition and language processing (Appendix A1.2.2). By contrast, attempts to substitute AI directly for human judgment in high-stakes or value-laden tasks raise more substantial methodological, ethical, and governance concerns, particularly when systems exhibit non-deterministic behavior or limited explainability (Appendix A1.2.4).

Beyond survey research, AI's rapid adoption is motivated by perceived benefits across multiple levels of society (Appendix A1.3). At a global scale, AI is used to support reasoning over large-scale, unstructured data in domains such as climate, energy, and public health, where it can act as a force multiplier for human decision-making (Appendix A1.3.1). At societal and civic levels, AI may improve information access, translation, personalization, large-scale verification, and the pace of scientific discovery through partial automation of research processes (Appendix A1.3.2). At the individual level, AI tools can support learning, accessibility, decision-making, and task management by adapting information to users' needs (Appendix A1.3.3). Economically, AI is widely framed as a general-purpose technology capable of raising productivity and reshaping

work through skill augmentation and selective automation, though realized gains remain uneven and context-dependent (Appendix A1.3.4).

At the same time, AI introduces well-documented risks that are especially salient for survey research. Generative models can produce fluent but incorrect or unsupported outputs (Appendix A1.4.1), complicating validation and reproducibility (Appendix A1.2.4). Their performance depends heavily on the quality and representativeness of training data, creating risks of bias, misrepresentation, or exclusion of under-measured populations (Appendix A1.4.1, A1.4.5). AI integration must also be evaluated against a broader set of concerns. These include misinformation and erosion of trust (Appendix A1.4.2), economic disruption (Appendix A1.4.3), model misalignment (Appendix A1.4.4), abuse and harassment (Appendix A1.4.6), environmental externalities (Appendix A1.4.7), and impacts on human cognition and well-being (Appendix A1.4.8).

Given the rapid pace of technological change, this report does not attempt to catalog all current or future AI tools. Instead, it emphasizes enduring principles that remain relevant across implementations: evaluating fitness for purpose, distinguishing augmentation from automation, and weighing productivity gains against risks to validity, transparency, and trust. Readers seeking more detailed discussions of AI models, capabilities, benefits, and risks, including technical terminology, adoption patterns, and broader societal implications, are directed to Appendix: Background on AI (Full).

Of these risks, the potential for AI to disrupt survey research as a profession warrants particular attention. Eloundou et al. (2023) identify survey research as among the most highly exposed fields to LLM-driven disruption, reflecting the close alignment between core survey tasks and the capabilities of generative AI. Disruption, however, is not synonymous with replacement. Recent analyses suggest that, for survey researchers, AI currently augments a larger share of job tasks than it fully automates, pointing toward a reconfiguration of work, often involving higher productivity and new responsibilities, rather than wholesale job loss (Washington Post, 2025). Crucially, the skills required to use AI responsibly are precisely those that define survey research expertise: designing unbiased questions, identifying error and bias, evaluating the validity of outputs, and assessing whether findings accurately represent the populations under study. Understanding which skills enable growth in the presence of AI, and how those skills should shape the integration, evaluation, and governance of AI in empirical research, will be central to the future of the field. The effects of AI on survey researchers and the survey industry are not predetermined; they will be shaped by the choices made today. We hope that reading this report and engaging in discussion with colleagues will help survey researchers shape the impact of AI in ways that strengthen the field and its profession.

3. AI's Use in Surveys

In this section we briefly explore not just how AI is currently being used in survey research, but provide an initial framework for understanding how AI may be used as the research and

technology continues to evolve over the next several years. Written in spring 2026, this report captures a moment in a rapidly evolving landscape, and the uses of AI in survey research will almost certainly extend beyond what we can reasonably anticipate here. §4 complements this discussion by introducing evaluation techniques, while §5 focuses on transparency and reporting recommendations.

The general survey workflow involves a number of steps that can be described with varying levels of specificity. A detailed outline includes discrete tasks such as defining study concepts through literature review; procuring vendors; designing samples; drafting and adapting data collection instruments such as questionnaires, recruitment screeners, and qualitative protocols; translating instruments; engineering survey software; recruiting and training interviewers; pretesting; targeting respondents; collecting data; harmonizing files; applying statistical adjustments; coding and thematically organizing qualitative data; and generating reports. A more stylized representation groups these activities into four broad steps or clusters of tasks: pre–data collection (where ideation around the study and survey occur), data collection (where the study is administered and fielded), post–data collection (where data are processed and analyzed), and briefing (where a report is created, disseminated, and consumed) (Rothschild et al. 2025).

Across each of these steps, AI has already begun to augment human work, most visibly through generative models such as LLMs. Researchers now use AI systems to refine research questions and hypotheses, iterate over operationalizations, and generate candidate survey items. AI agents are increasingly deployed during administration, both as automated interviewers and as simulated respondents for methodological exploration. In analytics, AI tools support transcription, translation, cleaning, classification, extraction of open-ended responses, thematic analysis, and modeling. AI can also assist in drafting reports, producing visualizations, retrieving data, and enabling interactive summaries. While work is occurring at each phase of the survey research process, the lion's share of early work is primarily focused on applications within the data collection and post-data collection phases (Buskirk et al. 2025a).

These uses represent the first stage of AI-driven disruption in the survey process: AI systems augment existing tasks without fundamentally altering the structure of the workflow itself (Rothschild 2025). Augmentation is characteristic of early adoption, when AI substitutes for or enhances specific subtasks but does not yet reorganize the surrounding production process. Over the longer term, Eloundou et al. (2023) identify survey research as among the most highly exposed fields, driven by both the direct application of LLMs to core survey tasks and the diffusion of LLM-enabled software whose capabilities closely align with central survey research activities.

There are emerging signs of a second stage of disruption, in which organizations begin reallocating resources across tasks as AI increases efficiency in some components of the workflow. For example, faster questionnaire iteration or automated coding of open-ended responses may shift staff time toward other aspects of the process including: pretesting, respondent targeting, or data interpretation. These changes can meaningfully reshape the cost

structure and pacing of survey work, but they still operate within the traditional conceptual survey lifecycle.

At the time of this writing, the field has not yet entered the third stage of disruption. This stage would involve researchers fundamentally redefining the survey workflow to be AI-first, reimagining the sequence, ownership, and granularity of tasks around the capabilities and limitations of AI agents. Such a transition would involve structural redesign rather than incremental optimization: new forms of instrument creation, adaptive fieldwork, continuous data quality monitoring, hybrid human–AI interviewing, or dynamic reporting environments built around AI-native pipelines.

This section documents uses of AI in survey research, focusing on where early task reallocation is already occurring (current prevalence) and risk to research quality (see Table 1). Rather than organizing applications around the linear phases of the survey research lifecycle, we structure the discussion along a continuum of increasing risk: AI as a Colleague, AI as a Briefer, AI as an Analyst, AI in Data Collection, and, finally, AI as a Workflow (Rothschild et al. 2025). We place AI as a Workflow at the end because, although it represents the highest-risk configuration, it remains largely aspirational and is not yet widely deployed.

Lower-risk clusters of tasks, particularly AI as a Briefer and AI as a Colleague, are characterized by stable, inspectable outputs such as draft questionnaires, analytic plans, or written summaries. These artifacts can be directly reviewed, audited, and revised by human researchers. Risk increases as AI systems move closer to data generation and end-to-end execution. When AI is involved in analysis, data collection, or orchestration of the full workflow, the core artifacts are no longer easily inspectable: a reader cannot feasibly review raw respondent: AI interactions, reconstruct how an AI interviewer adapted across respondents, or independently regenerate outputs from first principles. In these cases, trust hinges less on artifact review and more on transparency about process, instrumentation, and controls, precisely because the underlying artifacts are not readily observable.

Table 1. Risk and current prevalence of AI by clusters of tasks across the current workflow.

Clusters of tasks	Risk	Current prevalence
AI in Data Collection	High	Medium
AI as an Analyst	High	High
AI as a Briefer	Manageable	Medium
AI as a Colleague	Manageable	High
AI as a Workflow	Very High	Low

3.1 AI in Data Collection

AI asks and/or answers questions.

AI systems can now serve as active participants in the administration of surveys, either by asking questions (AI as an Interviewer) or answering them (AI as a Respondent). These uses raise some of the most important methodological, ethical, and operational questions in contemporary survey research.

3.1.1 AI as an Interviewer

AI conducts interviews or administers surveys or qualitative interviews interactively.

Vignette: A voice- and text-enabled AI chatbot administers a set of questions, adapting follow-ups based on prior responses while maintaining neutrality and standardized delivery.

Survey researchers are beginning to replace static questionnaires with conversational, AI-driven interviewing interfaces, or using AI chatbots to conduct qualitative research with participants. Powered by LLMs these systems can adjust mode, phrasing, pace, and follow-up questions in real time (i.e., accelerating moving into mixed-mode, and dynamic wording and branching). The promise is substantial: more engaging interactions, reduced respondent fatigue, richer open-ended responses, and improved comprehension for complex instruments that are capable of digging deeper into specific follow-up questions for sub-populations. Early pilots suggest these interfaces may increase completion rates for cognitively demanding tasks (lowering non-response error) and allow respondents to interact in a modality that feels more natural (minimizing measurement error) (Barari et al., 2025).

Importantly, the level of methodological risk increases with the degree of agency afforded to the AI interviewer. Risk is relatively low when systems are limited to modality switching (e.g., text versus voice), moderate when they introduce translation or paraphrasing, and high when they are authorized to generate new questions, alter question wording, or influence survey logic and flow.

Overall, adaptive interviewing introduces a fundamental methodological tension: the same flexibility that makes AI interviewing appealing also threatens standardization and comparability. Even subtle variation in tone, emphasis, timing, or follow-up logic may produce differential measurement error across respondents. Researchers must therefore evaluate whether LLM-driven dialogues preserve neutrality, avoid leading prompts, and consistently convey the same conceptual meaning across interviews, as well as determine what new forms of paradata are required to monitor and evaluate interviewer behavior at scale.

Operational and ethical risks are also salient. Although AI interviewers resemble tasks historically performed by human interviewers in call centers, they differ in two critical ways: the absence of real-time human oversight and respondents' understanding of the varied use-case of

their data. AI interviewers may ask inappropriate or off-script questions, mishandle sensitive topics, fail to enforce skip logic, or inadvertently signal approval or disapproval through wording or tone. These risks are amplified in voice-based interfaces where affective cues are harder to control. More broadly, conversational agents heighten concerns around consent, privacy, transparency, and auditability. Moving from passive web forms to active dialogue requires new norms for informing respondents about how their data are processed, stored, and used to adapt conversations in real time. Without such safeguards, perceived violations of trust may provoke backlash that ultimately degrades participation and data quality.

Looking ahead, this area will require technical safeguards, new validity frameworks, and research on respondent perceptions. This includes real-time logs, rule-bounded conversation engines, and audit trails. Research on respondent perceptions should address trust, comfort, and disclosure effects when the interviewer is an AI system. As conversational surveys expand, the challenge will be balancing innovation with the core methodological principles of standardization, comparability, and transparency.

Selected Examples from the Current Literature

A growing body of research evaluates the feasibility, quality, and methodological implications of deploying LLMs as AI interviewers. Wuttke et al. (2024) conduct one of the first controlled comparisons of AI-led versus human-led conversational interviews, finding that LLM interviewers can achieve levels of adherence to interviewing guidelines and response quality that are comparable to human interviewers while offering major scalability advantages. Complementing this work, Beltoft, Schneider-Kamp, and Askegaard (2025) study an “Interview Bot” designed for qualitative and ethnographic interviewing, showing that LLM-based agents can elicit meaningful, open-ended responses, though they still fall short of human interviewers in building rapport and executing nuanced follow-ups essential for deep qualitative inquiry (although, we expect these norms to change rapidly with the adoption of this type of technology into more everyday life). Tirumala and colleagues (2025) examine the fitness for use of AI interviews compared to IVR technology in both quantitative and qualitative contexts and find that effectiveness of AI in the interviewing process may depend heavily on context, for example, when human interviewers are unavailable, nuanced emotional detection is not essential, or topics are highly sensitive and prone to social desirability bias.

Experimental evidence from Barari et al. (2025) further demonstrates that AI interviewers can improve open-ended response quality: in a randomized survey experiment with 1,200 participants, AI-generated conversational probes substantially increased response specificity and detail, even with minimal fine-tuning to the domain or question context.

At the same time, evaluations in more naturalistic settings highlight limitations and areas requiring methodological caution. Cuevas et al. (2025) deploy a set of conversational agents working in tandem to interview 399 participants and find that while AI interviewers can enhance engagement and lower respondent burden, they continue to lag behind human-facilitated interviews when dealing with sensitive questions, contextual complexity, or participant expectations about interviewer roles. Research on “modular conversational agents” for surveys

shows promise for building domain-aware interviewers capable of incorporating specialized knowledge bases, but it also underscores risks related to inconsistent probing, privacy and data-security concerns, and potential biases stemming from opaque model training corpora (Yun et al. 2024). An agent designed to be your best friend may be good at probing in some ways, but may also create bias as it differs for each respondent.

Taken together, this literature suggests that AI interviewers are a credible augmentation to existing survey-administration workflows. However, realizing their benefits at scale will require explicit constraints, transparent design choices, and new validation standards to ensure that gains in engagement and flexibility do not come at the expense of measurement integrity.

3.1.2 AI as a Respondent

AI simulates responses for testing surveys or modeling answers or replacing human subjects.

Vignette: Before fielding to humans, AI generates synthetic responses to stress-test survey logic, identify potential skip-pattern issues, and estimate completion time.

A second emerging use of AI in the administration of surveys is the deployment of LLMs as synthetic respondents. In this approach, LLMs are used to generate sets of responses to survey instruments, often conditioned on demographic, attitudinal, or contextual information, in order to approximate how members of a target population might answer. Unlike classical imputation or synthetic data methods based on explicit probabilistic models, these responses emerge from a combination of user prompting and patterns learned from massive but not necessarily representative training corpora, and post-training alignment procedures (Argyle et al. 2025).

Before evaluating synthetic respondents on technical grounds, it is essential to consider whether AI-generated responses should be treated as a form of public opinion at all. Survey respondents and research participants are human beings whose behaviors, attitudes, and cognitions are directly measured. In the pre-LLM era, when plausible nonhuman alternatives did not exist, terms such as “poll”, “survey”, and “public opinion” implicitly assumed that data collection drew on human responses. By contrast, a response generated or inferred by an AI system is a model-based approximation of what a person might say, not a direct observation of human expression. At best, synthetic responses may function as proxies; at worst, they may systematically misrepresent the populations they purport to simulate. The field has not yet reached consensus on when, if ever, AI-generated responses can stand in for human ones without fundamentally altering what is being measured. At a minimum, researchers should be transparent about and clearly distinguish between data derived from human respondents and data generated by AI systems. We examine these distinctions in greater depth, and provide concrete recommendations, in §5.

There is a wide range of applications, with correspondingly different levels of reward and risk. We highlight three: (1) pre-field diagnostic testing, (2) post-field augmentation/imputation, and

(3) synthetic data collection as a substitute for human respondents. Overall, synthetic responses are the most risky of the core tasks we consider.

First, researchers are increasingly using synthetic respondents as a survey pretest. Before fielding to humans, teams can prompt an LLM to complete the instrument repeatedly, optionally under different persona constraints, to detect broken skip patterns, ambiguous wording, unexpected branching, and excessive completion time. This can accelerate the identification of design flaws that would otherwise require cognitive interviews, pilot tests, or intensive manual review (limiting measurement error). Synthetic response runs can also reveal edge cases (e.g., combinations of answers that trigger unintended loops or missing paths) helping teams "stress-test" complex logic earlier in the workflow. Further, synthetic respondents may help determine sub-populations that warrant further examination with increased targeting in the human sampling.

Second, a more consequential use is generating synthetic responses to fill in missingness or to anticipate subgroup patterns. Conceptually, this echoes approaches such as multilevel regression with poststratification (MRP), or other classical imputations, which uses observed survey data and population margins to project subgroup estimates under explicit modeling assumptions (Wang et al. 2015). LLM-based synthetic responding differs in that the mechanism is not a transparent statistical model; it is a prompted generative process that may extrapolate beyond observed human data in opaque ways. The risk increases when synthetic respondents are used to populate demographic clusters that are sparse or absent in the human data, or when they are used to infer changes over time or across items without strong empirical constraints. These risks mirror well-known limitations of multiple imputation when missingness is not missing completely at random (MCAR), and may be amplified when synthetic responses are generated without strong empirical constraints. This has the potential to mitigate the impact of coverage and non-response errors, but also exacerbate process error.

Third, some vendors now market "samples" generated entirely from synthetic responses. This is particularly risky: absent a grounding dataset of real human respondents, outputs can resemble plausible survey data while failing to reflect the true distribution of attitudes, the structure of correlations, or the heterogeneity within and across subgroups. For this reason, we prefer using the term synthetic responses in this report, rather than synthetic samples: the method is not a sampling design, but an attempt to estimate what a specified set of respondents would say. Eliminating the human respondents is basically outside the bounds of total survey error, which is constructed around the human respondent.

The risks are profound, especially when synthetic responses are used as a full replacement for human data collection. LLM-generated responses can sometimes approximate the marginal distributions of individual items, but they frequently fail to capture the nuance, contextual grounding, and heterogeneity present in real respondents (Bisbee et al., 2024; Morris et al., 2025; Wang et al., 2025). This is especially true for marginalized groups, culturally specific concepts, and emotionally charged topics, where publicly available models are particularly limited because they are fine-tuned to avoid controversial positions. Synthetic responses may

also create a false sense of representativeness, smoothing over real-world variability or amplifying existing model biases, thereby subtly shaping instrument design or interpretation in ways that erode validity.

As a result, the evaluation frontier is shifting from the narrow question of whether an LLM can match the average response to a single item toward the broader challenge of whether synthetic responses can reproduce relationships in the data. Future assessments will increasingly examine whether synthetic responses preserve correlations, joint distributions, subgroup interactions, and multivariate structure, not merely marginal frequencies. Earlier work on evaluating nonprobability samples (e.g., Kennedy et al. 2016) offers a useful conceptual foundation here, emphasizing validity at both descriptive and relational levels.

Online panels remain relatively inexpensive, complicating the economics of substituting synthetic responses for human participants. Yet as LLM inference costs fall below those of low-cost online completes, some researchers may tolerate higher error in exchange for affordability (at least for populations that are expensive but have plausibly have synthetics with higher accuracy, although there is likely a negative correlation between the expense of reaching a population and the accuracy of synthetic responses for that population). At the same time, tools for generating synthetic data are likely to become far more accessible, particularly as models are fine-tuned for specific domains, question types, or population subgroups. These refinements may help researchers stress-test instruments, explore edge cases, identify design flaws earlier in the workflow, or augment human samples with synthetic responses from hard-to-reach subpopulations. The challenge will be ensuring that increased ease of access does not incentivize overreliance: even as synthetic responses become cheaper and more capable, transparent reporting and careful validation will be essential to prevent synthetic data from subtly influencing real-world inferences.

Finally, it is important to note that “AI as respondent” has a second meaning: human respondents may themselves use AI. Respondents might (a) use LLM-assistance to lightly edit open-ended responses (analogous to spellcheck), (b) use LLM-augmentation to delegate writing to an AI tool, (c) use AI-automated agents as bots to complete surveys deceptively, or (d) use “digital clones” trained on their own past writing (either with or without validation). The remediation landscape here is fluid. Some uses (such as LLM-assistance or augmentation) can be addressed through instructions and disclosure, while deceptive use calls for detection, process controls, and transparency.

Selected Examples from the Current Literature

We note that the models used to generate synthetic responses are continually retrained, updated, or realigned without notice in ways that can alter their behavior over time in ways that are difficult to observe or document. As a result, conclusions about the congruence of synthetic responses depend on shifting human baselines and model performance in ways that may quickly and dramatically vary. Noting that results could change quickly as the AI technologies evolve, at the time of writing this report, the work on using synthetic responses in the survey research process is rather mixed. Some examples show that synthetic respondents can emulate

their human sample counterparts well while others paint a picture that is less optimistic. We present examples that highlight the continuum of what is currently understood about their use in survey research.

Argyle and colleagues (2023) introduced the concept of "silicon samples" (which we refer to as synthetic respondents) and criteria for assessing "algorithmic fidelity" for LLMs and reported nuanced similarities between human and AI-generated respondents. Bisbee and colleagues (2023) report contrary findings that suggest that synthetic respondents generate responses that are far less variable compared to actual survey respondents' responses. They also remark that results can be highly dependent on prompt and LLM versions being used. von der Heyde and colleagues (2025a) use ChatGPT 3.5 to investigate whether LLMs can estimate vote choice in Germany using the 2017 German Longitudinal Election Study. Estimates generated from the generated synthetic respondents exhibited a bias towards the Green and Left parties. The exhibiting of bias in liberal viewpoints has been one of the prevailing findings of many current studies employing synthetic respondents across various LLMs and countries (Santurkar et al., 2023, Anthis et al. 2025). Wang and colleagues (2025) found that synthetic samples can exhibit less variation than comparable human data thereby underrepresenting heterogeneity. Of course, there is a lot of available data in predicting vote choices, which is not available for other marketing or general survey topics.

An emerging body of work is beginning to examine how to better prompt and define synthetic respondents that go beyond direct specifications of a select number of socio-demographic and socio-political attributes. Park and colleagues (2024) compared these direct specifications to more comprehensive specifications that were based on transcripts from semi-structured interviews of survey respondents. These interview-based definitions of synthetic respondents outperformed those from direct specification on survey item accuracy and on personality inventories. Moon and colleagues (2024) also leverage open-ended backstories that were entirely LLM generated and match personas defined by these so-called "anthologies" to human respondents. Their process shows promising improvements in matching the response distribution of human respondents compared to other direct persona definition methods across multiple open-sourced large language models.

3.2 AI as an Analyst

AI processes and interprets data post-collection.

AI systems are increasingly serving as active participants in the post-collection processing of survey data. These applications include transcription and translation, data cleaning, labeling or annotation, and statistical modeling or estimation. While this cluster of tasks is generally considered high risk, the level of risk varies substantially by task. Uses that involve labeling, annotation, modeling, or estimation raise the greatest methodological concern, as they directly shape analytic decisions and inference rather than merely facilitating data preparation.

This is particularly notable given the speed and scale of adoption. In contrast to synthetic responses, where researchers have often voiced immediate and intuitive concerns about validity, AI-driven post-processing tools have been integrated rapidly and with comparatively little scrutiny. As a result, systems that influence categorization, measurement, and estimation are now widely used, often without the same level of methodological caution, transparency, or validation that would traditionally accompany changes at this stage of the survey workflow.

3.2.1 AI as a Transcriber and Translator

Vignette: Human conducts a qualitative interview or focus group and AI produces an instant transcription of the session.

AI has rapidly transformed transcription and translation by offering speed, scalability, and accessibility that were previously unattainable through human-only workflows. In qualitative research, particularly when working with interviews, focus groups, or open-ended responses, AI systems can convert spoken language into text with impressive accuracy and deliver instant multilingual translation. These capabilities sharply reduce time and cost barriers, enabling researchers to analyze larger and more diverse datasets without relying on manual transcription or translation services. In global studies, AI tools help bridge linguistic divides by supporting real-time interpretation, allowing researchers to include voices from varied cultural and geographic contexts more quickly and equitably. Collectively, these advances open the door to richer, more inclusive research and faster decision-making.

Yet significant concerns remain. Key questions include whether AI systems can match or exceed human transcribers and translators in capturing nuance, idioms, register, and cultural meaning, especially in low-resource or underrepresented languages. Errors in dialect recognition, speaker identification, or cultural context can distort qualitative insights in ways that are difficult to detect. This could easily reduce processing error but also add to it. These challenges raise fundamental questions about where the bottlenecks lie: Are they limitations of the underlying model architecture? Data scarcity? Prompt design? Computational constraints? Or deeper issues related to the representation of minority languages and cultural concepts in training data?

Future research must examine not only accuracy but adaptability. One particularly important frontier is whether real-time transcription and translation can support dynamic, branching conversations in which follow-up questions depend on participant responses. This capability is especially relevant for qualitative research and for work in the Global South, where linguistic diversity and cultural subtleties are central. Ultimately, the goal is not only to produce verbatim transcripts but to deliver context-aware, meaning-preserving, and culturally sensitive interpretations across languages. Whether AI can consistently achieve this level of fidelity, at scale and across contexts, remains an open and pressing research question.

Selected Examples from the Current Literature

AI-supported transcription and translation is becoming increasingly central to qualitative research workflows, with recent studies documenting their particular benefit, along with methodological and infrastructure challenges. Research on intelligent speech-recognition demonstrates that AI can substantially reduce both time and labor in converting audio into text while maintaining high levels of accuracy (Eftekhari, 2024; Mojadeddi and Rosenberg, 2024). The systems these papers explore offer multilingual capabilities and support large-scale data processing, making them especially valuable for global qualitative research. At the same time the authors note that the transcription and translation services require human review to detect errors involving accents, idiomatic language, and overlapping speech: areas where AI continues to struggle. Their errors will be most pronounced where there is linguistic diversity, cultural nuance, and low-resourced languages.

Beyond transcription, a growing body of literature examines the broader integration of AI into qualitative analysis, highlighting issues of interpretation, reflexivity, and epistemic validity. Reviews and conceptual frameworks suggest that AI can serve not just as an ex-post tool but as a potential analytical partner, reshaping coding and thematic analysis workflows (Bryda et al. 2025). However, critical scholarship warns that reliance on AI may risk flattening cultural meaning and obscuring context, especially in studies involving participants from the Global South or linguistically diverse communities (Eftekhari 2024). Together, this emerging literature underscores both the promise and the unresolved challenges of using AI for real-time transcription and translation in qualitative research, pointing toward the need for continued evaluation of accuracy, cultural sensitivity, and methodological integrity.

Moving from the qualitative to quantitative context, Tewari and Hosein (2024) conduct a proof of concept that shows how LLMs can be used to generate survey data sets using telephone survey transcripts using a series of zero-shot prompts. Their process produced a final data set that was 97% accurate across multiple choice survey questions.

3.2.2 AI as a Data Cleaner

Vignette: AI detects and corrects inconsistent entries, flags outliers, and imputes missing values using probabilistic models.

AI serves as a low-risk, medium-reward tool for cleaning survey data, particularly because survey data is inherently structured. This structure reduces the likelihood of major errors during automated cleaning, making AI a relatively safe (but lower-reward) choice. The benefits are real, if modest. AI can efficiently screen out ineligible respondents and flag data inconsistencies such as mismatched formats, missing values, or contradictory responses. These tasks, though routine, are time-consuming and prone to human error, making AI a valuable assistant in streamlining the preprocessing pipeline.

However, the use of AI in more interpretive cleaning tasks introduces subtle risks, especially when dealing with open-text responses. Unlike structured fields, open-text data requires

contextual understanding to identify and correct inconsistencies. In these cases, AI may overreach, removing valid nuance or introducing undetected errors. These risks are compounded when researchers rely on AI to make judgment calls without sufficient oversight. Thus, while AI offers medium rewards in structured survey contexts, its application must be carefully bounded to avoid undermining data integrity through overly aggressive or imprecise cleaning.

Selected Examples from the Current Literature

AI is increasingly used as a data-cleaning assistant in survey research, offering reliable performance on structured datasets where errors can be detected and corrected with relatively low risk (Ilyas and Rekatsinas 2022). Much of the literature is focused on more general machine learning, not on the most recent advances in LLMs, but systematic reviews show that machine learning in general is very strong at identifying inconsistencies, flagging outliners, and harmonizing variables formats, reducing human labor and improving downstream analytical performance (Côté et al. 2024). However, researchers have cautioned how these benefits are more clouded in unstructured data; when AI is used for more interpretive cleaning, such as processing open-text responses, it may misclassify unusual but valid responses or remove meaningful nuance, raising concerns about over-automation and validity (Gweon and Schonlau 2024). Current work by Allamong and colleagues (2025) offers a glimpse into using LLMs directly in data correction tasks. Specifically they apply LLMs (i.e., GPT-4o) to perform spelling corrections for about 50,000 open-ended survey responses using different zero-shot prompts and found that ChatGPT identified between 85 and 90 percent of the spelling errors noted by human reviewers. This may be an example of misappropriation of technology, as simpler tools can already check spelling, where LLMs provide a higher marginal value in scrubbing PII or careful redaction that are much harder to automate.

3.2.3 AI as a Labeler/Annotator

Vignette: AI processes open-ended responses or qualitative data and categorizes them into thematic clusters.

Researchers are increasingly using LLMs to analyze open-ended survey responses in real time, enabling scalable summarization, categorization, thematic coding, and sentiment analysis (Buskirk et al. 2025a). This lowers the cost and complexity of including open-text items, which were often avoided due to manual coding burdens. Some survey platforms now allow respondents to speak or type responses, with LLMs transcribing and interpreting them instantly, making mixed-mode surveys more feasible. These tools are already being used to extract political leanings, detect narrative presence, and measure sentiment toward specific topics or agents, often from the same set of open responses. The flexibility is powerful, and it’s reshaping how surveys are designed and analyzed.

However, this flexibility introduces risk. Small changes in prompts or model versions can lead to inconsistent labeling, raising concerns about measurement and processing error (this can be reduced by a machine-readable classification guide, a codebook, with a step-by-step process

into the system). Researchers must be cautious when mining the same open-ended responses to answer multiple distinct research questions, as this can amplify error and misinterpretation. Ethical concerns also loom large: commercial LLMs may store or repurpose sensitive data, and researchers often lack visibility into how these systems handle input. As use expands, the field will need standards for evaluating label quality, documenting model behavior, and ensuring that AI-generated insights are both valid and transparent.

Selected Examples from the Current Literature

According to the recent review by Buskirk et al. (2025a) one of the most common uses of LLMs within the post–data collection phase of the survey research process involves processing responses to open-ended survey questions. Mellon and colleagues (2024) compare the accuracy of six LLMs using a few-shot approach along with three supervised learning models to a set of human coders for categorizing the most important issue responses into 50 possible categories using data collected from an internet survey panel in England. Their work indicated that the highest achieving LLM's accuracy was within 1 percentage point of human coders and far surpassed that of any of the machine learning models. Comparable results have been replicated across a range of political science datasets and substantive contexts (Rytting et al. 2023). In a study aiming to classify discussions posted on two popular Chinese social media sites about the Russia–Ukraine war, Rogers and Zhang (2024) find general agreement between LLM and human classifications but mention that LLM grouped and classified more content as neutral compared to human reviewers: an outcome that authors note could be related to the HHH (helpful, honest, and harmless) principle that guides many public-facing LLMs. von der Heyde and colleagues (2025b) examine how different LLMs can be used to code open-ended survey responses in other contexts, using German data on reasons for survey participation. In their work, which compared several state-of-the-art LLMs of the GPT, Llama, and Mistral families using several prompting approaches, found notable differences among the models and noted that only a fine-tuned LLM achieved satisfactory levels of predictive performance compared to human coders.

Apart from coding open-ended responses, large language models are also being applied to other forms of public opinion (e.g., social media posts and comments to news articles) to perform annotations around stance, sentiment, or other similar features. Gilardi et al. (2023) compare annotations of LLMs to human annotators for a corpus of social media posts and news articles and found that ChatGPT's zero-shot accuracy exceeded that of human annotators by about 25% (relative to gold-standard created by trained research assistants) and a cost that was about 30 times cheaper.

3.2.4 AI as a Modeler or Estimator

Vignette: AI builds predictive models to estimate population parameters, incorporating weighting and uncertainty quantification.

Researchers are increasingly using LLMs to automate post-survey data processing tasks, where errors from manual coding, inconsistent transformations, and undocumented logic often

creep in. These models are already helping with variable recoding, logic checks, metadata generation, and even simulating pilot datasets to test workflows before full deployment. The appeal is clear: LLMs can streamline complex multi-stage processes with an agentic workflow, reduce human error, and improve reproducibility. As survey instruments grow more intricate, these tools are becoming part of the standard toolkit for cleaning and preparing data.

But automation brings new or heightened risks. Blackbox modeling is not new, but LLMs greatly expand opaque transformations that are hard to trace or validate, especially concerning that their outputs appear polished and authoritative. When constructing an agentic workflow, researchers must contend with limits in input and output windows that require dividing up the tasks, management of memory, and context into increasingly complex levels of automation and autonomy. If researchers rely on model-generated classifications or summaries without clear insight and documentation into these they risk embedding errors that distort downstream analysis. To mitigate this, best practices are emerging: maintaining audit trails of prompts and outputs, validating model decisions with human review, and integrating LLMs into human-in-the-loop systems. Most critical is making explicit where human judgment enters the process and what decisions are delegated to the system. Used carefully, these tools can reduce processing error and improve efficiency, but without transparency, they may become a new source of error themselves.

Selected Examples from the Current Literature

As we have seen in other examples, sometimes models perform better when they are fine tuned. However, fine tuning requires large enough data sets that represent the diversity in examples covering the categories one wishes to use in classification, for example. Ehrett and colleagues (2024) use data from a small health survey of hospital staff as examples in few-shot prompts provided to LLMs to generate additional synthetic survey responses. These synthetic responses were then combined with the responses from the small survey as the basis to fine tune another LLM for classification tasks. The fine–tuning approach results in an area under the curve value of 0.87 (i.e., a fairly accurate model).

Buskirk and colleagues (2025d) utilize zero-shot prompting approaches to directly generate estimates of a battery of financial outcomes using various versions of ChatGPT across an array of subpopulations. Their work indicated that temperature/reasoning settings can impact the accuracy of these estimates compared to estimates derived from human benchmark surveys and that the accuracy can also vary by the subpopulation being studies. As we discuss in §5, models and settings affect results, sometimes in predictable ways other times in unpredictable ways, but that is why it is critical to document and disclose choices.

3.2.5 Summary of Key Benefits and Risks

Across transcription, cleaning, labeling, and modeling, AI is becoming an integral analytical layer between raw data and substantive inference. Its primary benefits are speed, scale, and consistency: AI dramatically lowers the marginal cost of processing unstructured and multilingual data, making open-ended responses, real-time transcription, and large-scale

qualitative analysis routine rather than exceptional. In structured settings, AI can reduce clerical error, standardize transformations, and improve reproducibility; in unstructured contexts, it enables analytical workflows that were previously infeasible due to time or labor constraints. Collectively, these capabilities expand the design space of surveys and qualitative research, allowing richer data collection and faster analytical iteration.

At the same time, the risks AI introduces are systematic rather than incidental. Small, often opaque choices (e.g., model selection, prompt design, parameter settings, or fine-tuning data) can produce meaningful differences in outputs, creating new forms of processing error that are difficult to detect or audit after the fact. These risks are most acute in interpretive tasks, where AI may flatten nuance, misrepresent minority perspectives, or default toward neutrality in culturally or politically sensitive domains. The central tension is thus between efficiency and epistemic control: while AI reduces human labor, it also shifts judgment into less visible, model-mediated processes. Responsible use therefore depends not on adoption alone, but on integration: clear documentation, human-in-the-loop validation, and explicit disclosure of where decisions are delegated to AI. Used carefully, AI can reduce error and expand capacity. Used opaquely, it risks becoming a new and poorly understood source of bias.

3.3 AI as a Briefer

AI synthesizes findings for stakeholders.

AI systems increasingly function as briefers: synthesizing analytical outputs into narratives, visuals, and explanations tailored to stakeholder needs.

3.3.1 AI as Report Creator

Vignette: AI generates an executive summary with charts, key insights, and methodological notes, tailored to different audiences.

This is among the most mature and widely deployed applications of AI in research workflows. LLM-based tools are now embedded across content-creation environments (e.g., word processing, presentation software, dashboards, and even audio-visual reporting) making automated report generation increasingly ubiquitous. In survey research contexts, these systems typically operate downstream of analysis, translating structured outputs (tables, figures, model results) into natural-language summaries that foreground salient results while abstracting away technical detail.

Selected Examples from the Current Literature

Recent scholarship documents a clear shift from static survey reports toward AI-mediated briefing systems that translate analytical outputs into accessible, stakeholder-ready narratives. LLMs are increasingly used to generate executive summaries, highlight key findings, and contextualize results for audiences with varying levels of technical expertise. Empirical studies

show that LLM-based report generation can match or exceed human-written analyses in coherence and coverage when paired with structured data access and validation pipelines (Xu et al. 2025). Applied work in high-stakes domains such as clinical research further demonstrates that integrating LLMs with retrieval mechanisms can substantially reduce reporting time while improving factual consistency and traceability (Kuo et al. 2025). Collectively, this literature suggests that AI-driven briefing is rapidly becoming a core layer of analytical infrastructure rather than a peripheral productivity aid.

3.3.2 AI as Interactive Retrieval Experience

Vignette: AI powers a Retrieval-Augmented Generation (RAG) dashboard where users can query survey findings in natural language and receive context-rich answers with source citations.

Survey reporting is beginning to shift from static presentations and PDFs toward interactive, AI-powered tools that allow users to query data in natural language. Researchers and organizations are experimenting with chatbots and automated briefers that generate plain-language summaries, visualizations, and comparisons across survey waves. These systems can also integrate external datasets and translate findings across languages, making survey results more accessible to global audiences and non-expert stakeholders. The goal is to make data engagement more dynamic, flexible, and user-driven, especially for decision-makers who need quick, interpretable insights.

Selected Examples from the Current Literature

RAG systems allow users to ask natural-language questions of survey data and receive grounded, citation-linked responses, mitigating hallucination risks associated with standalone generative models. Research on RAG over structured and tabular data shows particular promise for longitudinal surveys, enabling multi-table reasoning and evidence-backed explanations on demand (Soliman and Gurevych, 2025). Nonetheless, the literature emphasizes persistent risks: AI-generated summaries can surface spurious patterns if retrieval or prompting is poorly constrained, and non-expert users may over-interpret plausible-sounding output. Accordingly, scholars highlight the need for guardrails such as retrieval validation, transparent citation surfacing, and user education to balance accessibility with analytical rigor.

3.3.3 Summary of AI as a Briefer

AI as a Briefer offers substantial productivity gains while introducing important but manageable risks. By translating analytical outputs into polished narratives, these systems can also amplify weaknesses in underlying data or documentation: obscuring uncertainty, data quality limitations, or provenance in a classic “garbage-in, garbage-out” dynamic. Poorly constrained retrieval or prompting may surface spurious patterns, and fluent natural-language output can invite over-interpretation by non-expert users. In response, contemporary implementations increasingly rely on guardrails such as retrieval validation, explicit citation surfacing, prompt constraints, and user-facing guidance designed to support data literacy. The central design

challenge is thus not whether AI can brief effectively, but how to balance accessibility with analytical rigor, ensuring that AI briefers expand stakeholder engagement without oversimplifying or distorting the underlying evidence.

3.4 AI as a Colleague

AI assists researchers in pre-data collection tasks such as constructing the research questions, and refining, crafting, or adapting those research questions into survey instruments.

There is a growing body of work showcasing the use of LLMs as collaborative partners in the earliest stages of survey work. These tools function like an extra colleague that helps researchers articulate study goals and develop question frameworks (reducing specification error); these tools can also generate draft items, refine wording, and adapt instruments across languages and modes (reducing measurement error). Because these tasks occur upstream in the survey lifecycle, where a clear and tractable artifact can easily be examined by an expert (i.e., a researcher can easily review the survey before it is launched), the risks are comparatively low and the potential productivity gains are substantial. The order below of writing questions, editing questions, and translating questions is in descending order depending on how much of the researcher's agency is offloaded to the LLM.

3.4.1 AI as a Question Writer

Vignette: A researcher provides a study objective (e.g., "measure trust in institutions"), and an AI model generates a menu of candidate questions, scale formats, and domain-specific probes aligned with standard guidance on clarity and construct validity.

LLMs are now routinely used to translate conceptual prompts into draft survey items. When given a construct definition, an AI system can propose question batteries, alternative phrasings, and different measurement strategies (e.g., Likert scales, forced-choice formats). This capability directly assists with reducing specification error by helping ensure that the operationalization matches the underlying theoretical construct. But, unless the researcher actively guides the process, if the researcher just accepts the survey as "good enough," this takes a lot of agency from the researcher, giving the LLM a lot of discretion.

Selected Examples from the Current Literature

Behrend and Landers (2025) walk through "effective prompt engineering, model selection, alpha and beta testing, launching, and monitoring" for using AI as a Colleague in the survey ideation process. Rothschild et al. (2024) explores the opportunity and risks of AI-generated questions in practice. For a popular audience, Shedlock (2025) details for Greenbook the latest AI tools for survey creation (both survey-specific and general). Buskirk and colleagues (2025b; 2025c)) explore how prompts can be optimized for generating survey questions via LLMs and how reading levels of generated questions can be steered using appropriate zero-shot prompting (i.e., with instructions and description of what to do, but without any task-specific examples) and

quality can be increased further with multiple-shot or ensemble methods. Their work indicates that without steering the reading levels of generated questions using an LLM model may be higher than desired for general survey purposes. Padgett and colleagues (2024) show that fine-tuning LLMs on example survey questions can be helpful for generating whole questionnaires, although their work notes that the question types generated can be influenced by the composition of the fine-tuning data set. Hernandez and Nie (2023) use about 3,300 items from the international personality item pool (IPIP) to fine tune an early version of ChatGPT then request the revised model to generate over 1 million personality items. Their evaluation showed these newly generated items were similar in length and were only slightly more difficult to read compared to the human generated items. Follow-up tests indicated that a sample of human reviewers could not distinguish human- and LLM-generated items with any regularity or significance.

3.4.2 AI as a Question Editor

Vignette: AI reviews draft questions for clarity, bias, reading level, and length, suggesting alternatives that may reduce measurement error.

Researchers use LLMs to critique existing items for potential pitfalls such as ambiguous references, unmatched time frames, or loaded wording. Some systems can also simulate how different demographic or cognitive profiles might interpret a question, offering insight into potential sources of differential comprehension (with the very serious caveat that LLMs have less understanding of minority sub-populations, potentially creating unintended leveling of their unique attributes). Emerging integrations within survey platforms allow AI to comment not only on text but also on interface elements that may influence measurement, such as response option layout or visual framing.

Selected Examples from the Current Literature

Yun and colleagues (2024) explored the application of large language models (LLMs) to generate varied versions of standardized questionnaires in an attempt to mitigate respondent fatigue while maintaining measurement properties of the alternate items within the context of repeated administration in longitudinal surveys. Barends and de Vries (2024) use instruction prompting with ChatGPT 4.0 to generate a shortened version of the HEXACO personality inventory. They also requested that the LLM make revisions to the generated scale to optimize scale properties including internal consistency and content validity. Their work shows that the LLM model's attempt at revising the initially generated scale did not produce significant improvements in these psychometric properties. But models have also improved dramatically in a short amount of time, so this type of result needs to be constantly updated, while also recognizing that not all model updates are better for all tasks.

3.4.3 AI as a Question Translator

Vignette: AI translates questions into multiple languages while preserving semantic equivalence, and flags cultural nuances for researcher review.

AI translation tools now support rapid multilingual instrument development. Beyond producing draft translations, some models can highlight idiomatic mismatches, cultural sensitivities, or terms lacking direct equivalents. This positions LLMs as useful aids for maintaining cross-linguistic comparability, though final adjudication still depends on expert review and cognitive testing.

Selected Examples from the Current Literature

Metheney and Yehle (2024) demonstrated that ChatGPT can meaningfully flag potential translation issues–such as ambiguous source wording, inconsistent conceptualization, inappropriate formality, and culture-specific concepts–when translating questionnaires into Castilian Spanish or Mandarin Chinese. Lee and colleagues (2025) find that the language of the prompt requesting translation can alter the formality of the translation and impute cultural context that may not be intended by the requestor. Their work also found considerable differences in the capabilities of different families of LLMs across a wide array of languages with DeepSeek (made in China) performing better translations of Chinese, but worse translations for Korean survey items. Models made and optimized for different languages and cultures will reflect that in their ability. Adhikari et al. (2025) report that respondents in the United States and South Africa perceived LLM-adapted survey items as clearer, more specific, and slightly less biased than conventionally translated items.

3.4.4 Summary of AI as a Colleague

The use of AI as a Colleague in survey design presents high-reward, manageable-risk scenarios. The primary advantages include:

- Acceleration of early-stage design, allowing researchers to iterate through more ideas and alternative phrasings.
- Improved alignment between constructs and measures, reducing specification error through rapid generation and evaluation of candidate items.
- Support for clarity and comprehensibility, particularly for novice researchers who benefit from structured feedback, reducing measurement error.
- Expanded multilingual capacity, allowing teams to test translations earlier and more efficiently.

The risks are real but comparatively contained. The most salient is overconfidence: non-experts may treat LLM feedback as authoritative, or misinterpret readability and quality scores without understanding their methodological limitations. Models may also produce superficially plausible but substantively flawed suggestions, requiring careful human verification. Transparency about training data, provenance, and model limitations remains essential, especially with our limited understanding of what survey elements are prevalent in the training data of closed large language models.

Looking ahead, progress will hinge on three things: survey-tuned models (because training data likely overrepresents public question wording and underrepresents codebooks, biasing models' views of "good" surveys), dual-mode tools for novices and experts (because novices need different outputs to build competence and produce quality instruments), and workflow-integrated interfaces that incorporate AI feedback into established methodological practice rather than replacing it.

3.5 AI as a Workflow

AI operates as an integrated pipeline, performing multiple survey tasks without clear boundaries or visibility.

Vignette: A researcher uploads a set of questions to a voice-enabled AI chatbot. The system administers the survey, clusters open-ended responses into themes, generates adaptive follow-up questions, and stress-tests them with synthetic respondents, then redeploys them to humans for several cycles. The researcher receives a final dataset and summary without insight into intermediate decisions.

> *Note:* Currently, such end-to-end AI-driven survey workflows are more common in academic exploration and early experimentation than in large-scale applied practice. As a result, the specific implementations described here are likely to evolve as new tools and capabilities emerge. Nevertheless, these exploratory systems are important to consider now, as they surface core issues of opacity, accountability, and methodological control that are likely to intensify as AI becomes more tightly integrated across the survey workflow.

Survey research has traditionally operated as a linear workflow: researchers define the question of interest, design the instrument, administer the survey, analyze the results, and brief stakeholders. This model assumes fixed questions, a single fielding period, and post hoc analysis. The integration of AI, especially in real-time flexible instrumentation, simulated responses, and analytics, introduces the possibility of a circular, adaptive workflow. In this emerging approach, researchers collaborate with AI to draft and refine questions, simulate likely responses to stress-test instruments, administer and monitor surveys in real time, analyze incoming data, and update wording, branching, and logic iteratively with human oversight. This creates a feedback loop that learns during fielding, allowing surveys to evolve dynamically while maintaining construct integrity. While adoption will vary across researchers, organizations, and use-cases, this model could provide faster, more responsive research that aligns closely with decision-making cycles. The AI workflow could involve a single agent conducting all parts of the process but more likely would involve multiple AI agents working in tandem or sequentially to perform various functions within the larger workflow. For example, there may be one AI agent created for conducting qualitative interviewing that is paired with a second AI agent that processes collected results from the first agent and suggests follow up questions and a third AI

agent that determines when to change topics of the interview as illustrated by the work of Cuevas et al. (2025).
The potential benefit of AI-first workflows (or workflows that incorporate multiple AI agents and human experts) could be significant. By reducing cost and cycle time, AI makes it possible for iterative testing and rapid deployment at scale. AI-driven response simulation can pretest instruments, anticipate distributional responses, and surface potential failure modes before launch, unlocking questions that were previously too costly or complex to answer. And humans can be injected in the loop for validation at each step from question design, the response, to analytical review. This could empower a transition from discrete waves to continuous sensing, where surveys function as living instruments that ingest responses, detect emerging themes, and refine questions midstream. This adaptability strengthens the connection between survey operations and market intelligence pipelines, embedding surveys into decision-making rather than treating them as standalone exercises.

However, the circular model introduces new risks that must be managed carefully. AI systems can be opaque, with outputs shifting due to model updates or subtle prompt changes, which complicates understanding and reproducibility. Measurement error may increase as there is less control over the exact details of the questions and logic. Adaptive instruments add non-stationarity risk: mid-field changes can undermine longitudinal comparability without strict guardrails. Complex failure modes arise from interdependent components, including models, prompts, and orchestration layers. This interdependence makes subtle bugs harder to detect and correct. Overreliance on simulated respondents can bias question wording toward model priors rather than real populations, skewing instruments toward patterns that reflect AI assumptions rather than actual respondent behavior. Ethical and compliance considerations, including consent, privacy, and transparency about AI involvement, also become more pressing as workflows evolve. These challenges require research, transparent tooling, and well-defined human-in-the-loop checkpoints for validation and learning.

There are many open research questions before this is viable. Under what conditions do AI-assisted workflows match or exceed human benchmarks for validity and reliability? Where does human intervention have the greatest marginal impact, and what is the minimum effective dose of oversight needed to maintain acceptable error bounds? How can adaptive instruments update questions mid-field without compromising construct validity or trend comparability? What protocols can detect and correct bias across demographic and linguistic subgroups? What combination of automated metrics, human adequacy ratings, and construct fidelity tests should define the standard evaluation framework? And what governance practices, including disclosures, versioning, and reproducibility standards, are necessary to ensure trust and compliance? These questions will shape the methodological foundation for AI-driven survey research.

To support this transformation, tooling must evolve. Systems need transparent translation and transcription interfaces with confidence scores and escalation paths for human review, question co-creation tools anchored to validated construct libraries with bias checks and style controls, and response simulation platforms that clearly separate projections from observed data while

monitoring drift. Orchestration for adaptive branching should include pre-registered update rules and freeze points to maintain comparability, while evaluation suites should bundle transcription and translation metrics, human ratings, and construct-consistency checks. Standard tooling must include immutable logs, replay capabilities, and version tracking, along with integrated privacy and compliance controls. Human-in-the-loop processes should be designed for efficiency, with role-based queues and escalation paths that fit real team structures. Finally, integration with analytics pipelines should allow validated signals to stream into business intelligence systems while preventing premature automation of decisions based on unverified adaptive changes.

4. Evaluating the Usage of AI for Survey Research

This chapter presents a broad framework for evaluating data quality and error, as well as the outputs and quantities of interest (e.g., estimates, effects) computed when using artificial intelligence, specifically generative AI (GenAI) models such as large language models (LLMs), vision language models (VLMs), multi-model language models (MLLMs), in survey and social science research. GenAI tools could be used either explicitly or implicitly in third-party software and tools, and we recommend performing evaluation in either case. The framework is followed by illustrative examples demonstrating how it can be applied to different use cases. In practice, the framework should be tailored to the specific application under consideration. §3 complements this discussion by detailing how AI is used in surveys, and §5 provides transparency and reporting recommendations.

A key premise of this framework is that GenAI must be evaluated in relation to the *specific task(s)* they are intended to perform. Prior evidence that a GenAI system performs well in a related or superficially similar task does not guarantee comparable performance in a new context, as change in context (e.g., domain, subpopulations, language) can introduce distribution shifts. Similarly, when orchestrating a series of tasks in a multi-stage workflow, a locally optimal GenAI model may not perform as strongly as it does on the single task it was originally evaluated for. This issue is discussed in greater detail later in this section.

At a high level, the evaluation framework focuses on four core criteria related to:

- **Validity:** Does the GenAI tool target the intended task, rather than an irrelevant or superficially related but divergent task?[2]
- **Performance:** How effectively does the GenAI tool perform the intended task (according to relevant measures to the task, e.g., accuracy, variance)?

[2] Validity here shares a nearly identical meaning to *construct validity*. However, since LLM procedures for survey research may not always be used explicitly for measurement, we use a more general term referring to “task validity.”

- **Sensitivity:** Do the results of the intended task significantly change when the process, prompts[3], data inputs, or models are slightly varied? Do different GenAIs produce similar results for the same task with the same instruction, input and/or data?
- **Reliability:** Does the system that incorporates GenAI tool(s) produce consistent substantive results when repeated under the same conditions? Here, consistency refers to the reproduction of estimates, effects, or conclusions, rather than replication of the exact free-form text.

This framework is conceptually similar to established evaluation paradigms in survey research, such as total survey error and the fit-for-purpose framework (Groves et al. 2009, Statistics Canada 2017, Tabassi and NIST 2023). As with those approaches, the goal is not to define universal benchmarks or thresholds for acceptable performance. Instead, the framework provides a structured checklist of considerations to guide researchers in evaluating LLMs across multiple dimensions of quality.

4.1 A Few Do's and Don'ts

Before detailing the evaluation criteria, we outline a set of practical do's and don'ts intended to support responsible and transparent use of GenAI in research. These recommendations complement the evaluation framework and reflect common challenges encountered when integrating GenAI into applied research workflows. Following these guidelines can improve transparency, facilitate replication, and support meaningful evaluation of sensitivity and reliability.

4.1.1 Do involve humans in all GenAI-mediated survey research activities

There are several known threats to the validity of GenAI in research activities. GenAI models are known to hallucinate content (e.g., LLMs produce statements that are fluent but factually incorrect) (Guerreiro et al. 2023, Alansari 2025). They may also misinterpret prompts and generate outputs that differ from what the researcher intended (NIST 2024, Bean et al. 2025). In addition, models may not consistently adhere to implicit or explicit guardrails unless these are carefully specified, tested, and monitored (Zou et al. 2023, Hakim et al. 2025).

For example, researchers may use LLMs to replace or supplement interviewer behavior in deconstructed or conversational interviewing, where follow-up questions are generated dynamically based on prior responses. In such settings, an LLM may begin with a standardized question written by the researcher but then generate follow-up questions that are inappropriate, irrelevant, or misaligned with the research objectives. Researchers should therefore test LLM behavior using a wide range of plausible responses and carefully review outputs to ensure the system remains within acceptable bounds. Soft launches, conducting a small number of cases and reviewing transcripts before full deployment, can be especially effective for identifying failure modes early.

[3] Prompt includes system prompts, templates, examples, chain-of-thought and other instructions that are provided as context to the LLM along with any system prompts intrinsic to the LLM model.

4.1.2 Do not assume that performance in one use case will generalize to another

Researchers should not assume that strong performance in a prior application will carry over to a new task. For instance, an LLM may translate English survey questionnaires into Spanish and Chinese with reasonable accuracy, yet perform substantially worse when translating into another language (Lee, Tian, and Morales 2025). As a result, a task-specific test dataset should be created and evaluated for each new application.

Differences in performance may arise not only from the substantive task (such as language or topic), but also from differences in prompts, model architecture, model version, parameter settings, or other aspects of system integration that may not be readily observable. For example, LLM performances are often not robust when there are distribution shifts between validation or calibration data and new data, and in the presence of distractors (Shi et al. 2023, Chen et al. 2025, Chen et. al. 2026). Each application therefore warrants its own evaluation.

4.1.3 Do not assume that past performance will generalize to future performance

Similar to the previous point, performance observed during initial testing or early deployment should not be assumed to remain stable over time. Unlike many traditional statistical tools, AI systems are often embedded in evolving technical and operational environments. Model providers may update underlying systems, prompt templates may be revised, or integration layers may change in ways that subtly alter model behavior. Even when the nominal model remains unchanged, shifts in the study population can affect how the system performs. Models that are connected to other sources of information (such as the Internet or external databases) may be especially prone to unanticipated changes in performance.

For this reason, evaluation should not be treated as a one-time pre-deployment activity. Instead, researchers should consider evaluation as an ongoing process that continues throughout the period in which an AI-enabled system is used in production. During this post-deployment phase, continued monitoring of evaluation criteria can help identify unexpected performance changes, shifts in output distributions, or differences in subgroup behavior that were not observed during initial evaluation (Rao et al. 2026).

4.1.4 Do document all decisions and procedures.

Replication, or at least the less rigorous reproducibility, is a foundational principle of social science research and is particularly important when using GenAI. Small changes in how GenAI is integrated into the research workflow, such as prompt wording,[4] preprocessing steps, or

[4] Prompt includes system prompts, templates, examples,chain-of-thought and other instructions that are provided to the LLM as well as intrinsic system prompts. Therefore it is important to note down the specific model and exact design of the prompt to be able conduct evaluation and for reproduction of results.

post-processing rules, can lead to meaningfully different results (Barrie et al. 2024, Palmer et al 2024).

Researchers are therefore strongly encouraged to document all substantive decisions and procedures, including model selection, model parameters, memory, method of calling model, prompt design, validation steps, and human oversight protocols. Detailed documentation supports evaluation of sensitivity and reliability, enables replication, and facilitates transparency for reviewers, collaborators, and other stakeholders. We explore disclosure in detail in §5.

4.1.5 Do not solely rely on LLMs to evaluate LLM-mediated survey research activities.

The inverse of the first recommendation is also true, particularly when it comes to evaluation. We caution against an evaluation strategy built on "LLMs all the way down," where the same class of systems generates outputs and then serves as the primary arbiter of their quality. This is particularly worth noting given the increasing popularity of the "LLM-as-judge" paradigm in AI evaluation research, where LLMs are used to score or compare model outputs (Zheng et al. 2023, Gu et al. 2024). While it has been shown that strong LLM judges are capable of achieving high agreement rates with human evaluators, a range of other biases have also surfaced with this practice, for instance the explicit preference that LLMs have for other LLM outputs (Laurito et al. 2025).

The appropriate degree of oversight will vary across applications, depending on the *risks* of the application, the *resources* available to the research team, and the *rectifiability* of errors before they affect substantive conclusions or respondent experience. In high-risk or high-consequence settings, such as tasks involving respondent interaction, sensitive populations, consequential inference, or limited opportunities for correction, more comprehensive human review may be warranted. In lower- or moderate-risk applications, structured oversight may take the form of targeted sampling, periodic auditing, or review at key decision points rather than line-by-line inspection of every output.

That said, for well-defined tasks with clear ground truth (e.g., accuracy of basic codes or adherence to a fixed coding rubric), AI can still be useful as evaluators of the performance dimension described above, provided a layer of human supervision (Gilardi et al. 2023) and statistical validation with human responses (Vishwakarma et al. 2023). When assessing other dimensions, particularly validity (the importance of which will be clear in the following section) we recommend stronger human oversight.

In all cases, responsibility for the work remains with the human researchers, not the system itself. Researchers should therefore make explicit who is responsible for oversight, what form that oversight takes, and how decisions about validation, escalation, and review were made.

4.2 Evaluation Criteria

As noted above, the evaluation criteria described here are intentionally broad. They are designed to guide researchers in assessing the relevance and quality of outputs produced by GenAI systems, rather than to prescribe a single correct method of evaluation. The appropriate implementation of each criterion will depend on the application, data source, and substantive goals.

Here we expand our four key evaluation criteria. Depending on the specific application, we note that other criteria may be essential and provide some examples of such in §4.4. We also emphasize that transparency should be treated as a "meta-criterion": it does not mainly assess the quality of the output itself, but whether the use of the system can be understood, interpreted, scrutinized, reproduced, and governed by others. For this reason, we treat it as a distinct category deserving of special attention in §5.

4.2.1 Validity

As used in this report, validity concerns whether the GenAI model is producing outputs that are appropriate for the intended task – akin to the social scientific definition of "construct validity," but applies to a more general class of tasks beyond measurement. Researchers should begin by asking whether the model's outputs meaningfully align with the intended concept and task.

There may be multiple valid approaches to assessing validity. For example, researchers may use LLMs to create digital twins of actual respondents to scale responses to new questions. Researchers may establish validity by assessing whether outputs target the correct construct–that is, they resemble realistic human responses from the target population, not merely surface-level plausibility, stereotypes, or artifacts of the model's training data. The evaluation may include checks for unrealistically high rates of refusal, selections of non–existent response options, implausible combinations of response options, or irrelevant outputs. When used to predict open-ended responses, researchers may include checks for training data leakage (e.g., unusually specific strings that may be reproduced verbatim from pre-trained sources).

Other applications may require different validity checks, such as expert evaluations, explicit model confirmation of the task instructions, or downstream impacts on other relevant tasks (similar to psychometric tests of "predictive validity"). The exact validity requirements are left to the discretion of the researcher and should be determined based on the goals and constraints of the specific study. However, validity is a necessary precondition for all GenAI applications. Absent such evidence, performance estimates lack clear interpretability, since they may capture apparent success on a proxy, artifact, or superficially similar task to the one at hand.

4.2.2 Performance

Once validity can be ascertained (the model *can* perform the task), researchers should measure whether the GenAI model can systematically perform the task *well* according to concrete

metrics relevant to the specific task, e.g., accuracy, variance.[5] Importantly, these evaluation metrics and the acceptable performance thresholds should be established prior to deploying the GenAI tools in production.

For example, if an LLM is used to code open-ended survey responses, researchers may statistically compare the distribution of LLM-generated codes to those produced by human coders on a shared subset of responses. Agreement metrics such as Cohen's kappa can be used to assess concordance between the LLM and human judgments on individual responses. Alternatively, researchers may treat human coding as a benchmark and evaluate the LLM's false positive and false negative rates. For prediction tasks, using separate validation data from human respondents drawn from the target population for the task at hand to evaluate performance is crucial. A GenAI tool might perform well in predicting responses to questions in a survey collected in the past, especially since the relevant contexts (or the data itself or results) could be part of the data used in training the GenAI model, but the same GenAI tool might not perform well in predicting future responses that depend on new current events.

For their particular application, researchers do not need to reinvent the wheel for measuring performance: a large and rich collection of metrics exist for effectively evaluating different aspects of LLM performance (Hastie et al. 2009, Naidu et al. 2023). The critical requirement is not the specific metric chosen, but that some explicit criteria for accuracy relevant to the task are defined, measured, and evaluated. GenAI should not be used in substantive analyses until the criteria of validity and performance are met, as this restriction helps ensure adequate data quality, results and interpretability.

4.2.3 Sensitivity

Sensitivity assesses whether results produced by a GenAI model are robust to reasonable changes in inputs, procedures, or model choice. Researchers should examine whether small or plausible variations lead to substantively different conclusions.

Sensitivity checks may include rewording prompts, altering input data, changing the population or context to which the GenAI is applied, or substituting one GenAI model for another. Prior research and applied experience suggest that relatively minor changes in prompts or system configuration can yield meaningfully different outputs, even when the task appears unchanged (Shi et al. 2023, He et al. 2024, Sclar et al. 2024). This sensitivity is especially consequential when demographic subgroups are small or sparsely represented: small variations in demographic cueing (names vs. explicit attributes; first-person vs. third-person framing) can shift the model's inferred persona and lead to meaningfully different outputs and performance

[5] We note that "performance" errors may map onto multiple stages of the total survey error (TSE) framework. Depending on the particular LLM application, errors in performance may encapsulate different TSE errors. For instance, if an LLM is used for open-ended response coding, performance error is equivalent to processing error, while if used to generate code to produce the final data file, performance errors may be described by analytic error in TSE parlance.

estimates (Tonneau et al. 2026, Weeber et al. 2026). As a result, sensitivity testing is a critical component of GenAI evaluation.

Conceptually, sensitivity testing parallels robustness checks commonly used in statistical analysis. Just as researchers assess whether results are stable across alternative model specifications, they should assess whether conclusions drawn from LLM-assisted processes remain consistent under reasonable variation. When results are highly sensitive to small changes, researchers should exercise caution in interpreting or generalizing findings. In some cases, it can suggest that the validity of the LLM task itself may be flawed.

4.2.4 Reliability

Reliability concerns whether the same GenAI-enabled process produces consistent substantive results when repeated under the same conditions. Unlike sensitivity, which focuses on variation in inputs or models, reliability focuses on stability over time and repetition.[6]

GenAI presents particular challenges for reliability assessment. They are often opaque systems with different sources of randomness arising in software and hardware implementations and their behavior may change as models are updated, retrained, or modified by providers (Barrie et al. 2024). As a result, even when researchers use identical inputs, prompts, and procedures, repeated runs may yield different outputs or lead to different downstream estimates.

For example, a researcher may use an LLM to classify survey responses into topical categories on a weekly basis as new data arrive. Even if the same prompt and model version are used, the distribution of classifications may drift over time, potentially altering trend estimates or subgroup comparisons. Without explicit monitoring, such changes may go unnoticed and be mistakenly interpreted as real changes in the underlying population.

Researchers who rely on LLMs over extended periods should therefore build in ongoing reliability checks. These may include periodically re-processing a fixed reference dataset, tracking key output distributions over time, or monitoring whether core estimates remain stable across repeated runs. Detecting inconsistencies early can help prevent gradual degradation in data quality and reduce the risk of incorrect substantive conclusions.

4.3 Examples of Evaluation

The above evaluation criteria are broad by design to apply to the diverse set of GenAI applications in survey research. However, the broad nature of them may be perceived as abstract, and it may be difficult for researchers to see how to apply these criteria to their specific situation. This section walks through a variety of GenAI applications and ways in which these criteria may be applied.

[6]See Rabanser et al. (2026) for a comprehensive framework for defining and evaluating reliability with the usage of AI agents.

While some applications are self-contained tasks that may involve only a small number of interactions with an LLM (e.g., editing a survey question), others are multi-stage workflows that may involve the usage of GenAI agents (e.g., conducting an entire analysis of a survey dataset). Our evaluation criteria apply to all possible usages of GenAI along this spectrum. We note, however, that this is not an exhaustive list of scenarios. Additionally, the example metrics used are not the only way to evaluate GenAI applications. They are meant as examples only.

4.3.1 AI as a Colleague

As we detail in §3.4, across writing, editing, and translating survey questions, LLMs are increasingly used as colleagues: productive partners that generate options, surface issues, and accelerate iteration, while humans make final decisions. The same evaluation dimensions apply across tasks. We begin by laying out some general recommendations for each dimension and then provide examples of more specific failures and successes across specific roles.

Validity. Even though the goal is not to perform a particular analytical or measurement procedure, researchers must confirm that LLM outputs are correctly targeting the intended task construct. Above all other evaluation criteria, it is essential that researchers establish the model can perform the role of a trusted, informed, and well-trained colleague. At a minimum, it should be clear that model outputs are relevant to the specific question or project, demonstrating that the LLM meaningfully comprehends the specific task.[7] Although this may not involve a systemic evaluation of "large-n" outputs per se, evaluating validity still requires–as it would with a human research assistant–expert review, testing, and feedback (e.g., adjustment of prompts). One prompt-based validity strategy involves probing the model to confirm the research objective before it generates any substantive outputs. For example, the researcher could prompt a model to (a) summarize the project goal in its own words, (b) list the key assumptions and scope boundaries, and/or (c) specify what would count as a "correct" output versus an out-of-scope response. Many similar techniques exist within a class of model prompting methods known as chain-of-thought prompting (Wei et al. 2022). In all cases, the researcher is still responsible for interpreting the results of this probing and ensuring that it reflects a relevant encoding of the task (the validity dimension) before moving onto a systematic evaluation of outputs (the performance dimension).

Performance. Although the key evaluation dimension in the AI-as-a-colleague usage model is validity, LLM outputs must still be judged systematically for their effectiveness, particularly if used repeatedly across many different tasks or projects. Human evaluation, through expert review, cognitive interviewing, pretesting, or similar methods, remains essential, as models often rate their own outputs favorably and may miss subtle interpretive problems (Laurito et al. 2025). Given the internal-facing nature of this usage pattern where researchers are the initial and often primary consumers of the model output, performance evaluation of AI-as-a-colleague

[7] In assessing validity, while it is important to demonstrate that the task is being performed correctly, researchers should avoid making anthropomorphic statements about "machine understanding" that implicitly equate LLM behavior with human understanding. This is particularly important in AI-as-colleague usage models where interactions may feel more collaborative or human-like.

should focus on researcher utility, rather than respondent utility. Procedures may include blinded expert ratings of output quality against a rubric (e.g., clarity, completeness) produced after each collaboration session or auditing output samples for error types (e.g., incorrect edits, unsupported claims, missed edge cases). Specific quantitative metrics may include average expert rating scores, percentage of outputs requiring revision, mean time saved per task, or research team satisfaction scores.

Sensitivity. While variation across prompts, models, and runs can pose a threat in analytical applications, it can be an asset when LLMs are used as collaborative tools. Here, we recommend that researchers treat LLMs as interactive idea generators: researchers can intentionally leverage this sensitivity to generate alternative framings, surface edge cases, and stress-test assumptions by varying prompts or model configurations. For example, researchers may pair open-ended prompts (e.g., “Tell me the best wording of question X”) with close-ended prompts (e.g., “Which of the following options is the best wording of question Y?”). They may also use open-ended prompts to elicit novel ideas or reasoning and then follow with close-ended items to standardize, validate, or compare those responses across respondents. When differences arise, it is the researcher’s discretion and responsibility to triangulate stronger wording or reveal where additional human judgment is required, rather than to pick a single “right” answer.

Reliability. LLMs may produce different recommendations across seemingly identical runs, similar to how inquiring a colleague may surface different insights at different points in time. As with human collaboration, researchers should be cautious in attributing specific reasons for differing recommendations. When the same LLM is used to surface recommendations across many tasks within a single session, later outputs may not be independent draws: they can be shaped by prior user prompts and the models’ own completions. In drafting and review, this variability can provide productive breadth. However, reliability becomes a concern when recommendations diverge in ways that would meaningfully change research decisions or downstream outputs. In such cases, researchers should trigger human safeguards–such as reinitialization of a “fresh” model session, targeted expert review, or querying the model to resolve diverging recommendations–so that variability is managed systematically rather than implicitly.

While we describe many aspects of the process that warrant attention, these tasks are distinctive in that they produce a concrete artifact–a full or partial survey instrument–that can (and should) still be straightforward to evaluate through human review. That said, the appropriate evaluation approach differs across the specific task of the AI-as-a-colleague. Table 1 below provides an overview of these considerations as well as actionable recommendations.

Table 2. Examples of survey research tasks (AI as a Colleague) and associated evaluation criteria.

Task	Core risks	Key evaluation recommendations
Question writer (§3.4.1) ● Generate ideas for new questions ● Propose alternative phrasings and operationalizations of a construct ● Suggesting synonymous terms appropriate for the target population	● Biases in generated questions ● Inconsistency in generated response scales ● Incoherence with other questions or project context	● Evaluate **performance** with human methods (expert review, cognitive interviews, pretests) ● Use **sensitivity** to your advantage by sampling multiple prompts/models to broaden options ● Treat variability across runs as useful diversity rather than a risk to **reliability** ● For accurate evaluation, maintain consistent prompting instructions across items
Question editor (§3.4.2) ● Flag and edit double-barreled, leading, or other biased questions ● Improve the comprehensibility of questions with alternative terms appropriate for the target population ● Correct flawed or mismatched response scales	● Failure to correct biases ● Failure to incorporate project context or target population ● Falsely identifying non-issues	● Calibrate **validity** early*:* compare LLM reviews with human expert reviews (prospectively or using previously annotated questionnaires) ● For sensitive or difficult topics, quantify **performance** where ground truth is available (e.g., false positives/negatives of biased terms) ● Leverage **sensitivity** by running multiple models or prompts to surface a wider set of potential issues ● Use human judgment to resolve discrepancies
Question translator (§3.4.3) ● Translate questions into multiple languages ● Flag cultural nuances (e.g., idiomatic mismatches) for researcher review	● Failure to preserve semantic equivalence between input and output language ● Failure to identify or translate regional nuances	● Human safeguards are crucial to bolster **performance** and **reliability**: bilingual review, comparison to prior approved translations, and back-translation where appropriate. ● Provide explicit detail in prompts to ensure **validity**: specify regional dialects (e.g., U.S. Spanish vs. Spain Spanish), terms that need special handling (e.g., domain jargon, culturally loaded terms, or concepts without settled translations) ● Use experts to quantify **performance** of translations, specifically in appropriateness for target population ● **Reliability** concerns are largely procedural rather than item-specific: prioritize consistent process application for new materials

4.3.2 AI as an Interviewer

As we detail in §3.1.1, LLMs may be used as interviewers in several ways. In more structured applications, similar to interactive voice response (IVR) systems, an LLM may read or display survey questions verbatim and answer respondent inquiries such as requests for clarification about the study sponsor or question intent. In more flexible implementations, LLMs may conduct conversational or cognitive interviews, where the researcher provides a baseline question and the model follows up as needed (e.g., “What were you thinking about when I asked you about X?”). In this role, the LLM functions as an interviewer colleague, shaping the interaction while operating within constraints set by the researcher.

Validity. This dimension is central when LLMs interface directly with respondents, ensuring that respondents are not misunderstood, confused, harmed, or experience other negative interactions. Rigorous testing should be conducted to ensure that any “hard” rules are applied consistently, such as reading questions verbatim, refraining from offering definitions, or avoiding inference when a respondent’s answer does not align with the available response options (e.g., mapping “most of the time” to “often”). Transcripts from testing or a soft launch should be reviewed to identify behavior that is out of bounds (e.g., off-topic follow-ups, inappropriate language, or the expression of opinions) or that results in unproductive loops (e.g., repeated prompting after “don’t know” responses). While it is crucial to preserve the respondent experience, validity also requires that–in the traditional meaning of construct validity–the correct underlying construct for a survey question or sequence of related questions is targeted across all respondents. We discuss this later in relation to reliability and sensitivity concerns.

Performance. The AI as interviewer usage model can often be decomposed into relatively mechanical subtasks that can be executed with more or less effectiveness and can be quantified systematically. For example, when interviews involve oral interaction, researchers can assess whether the system accurately transcribes respondents’ answers and whether any downstream coding or extraction steps reproduce what reliable human coders would produce. Similarly, identifying when “relevance probing” is warranted–because an answer is incomplete, ambiguous, or low-information–can be operationalized and evaluated using standard error metrics (e.g., accuracy, false positives, and false negatives). Based on the specific performance issues that surface, researchers may need to refine the configuration of the AI interviewer by adjusting prompts, iterating on model training or fine-tuning, revising decision logic, or adding and validating “hard rules” that constrain behavior in predictable ways.

Sensitivity. Considerations about sensitivity are especially salient because interviewer effects may vary across respondents. Researchers may wish to test how the LLM responds to different voices or accents in oral interviews, and to typos, slang, or emojis in written interactions. Different LLMs–or different configurations of the same model–may elicit different responses due to variation in speed, voice characteristics, sentence structure, or the degree to which the model adopts human-like conversational behaviors. Testing these features can help researchers identify interviewer behaviors that are appropriate for their population and research goals. In unstructured or semi-structured interviewing, different models or prompts may also generate

different follow-up questions; researchers may assess which configurations yield probes most consistent with human interviewers.

Reliability. Concerns about reliability arise primarily in unstructured or fully adaptive or conversational interviewing. LLMs may adapt based on prior interactions, potentially altering follow-up questions over time. LLMs may also produce different follow-up questions, which may only slightly differ in wording or have diverged in meaning completely, to the same respondent answers. In some contexts, such adaptation or diversification may be desirable (see Yun et al. 2023); in others, it may introduce unwanted variability. Researchers can monitor this by periodically running fixed test scenarios through the interview flow to assess whether the LLM's behavior remains consistent and within predefined bounds, adjusting prompts or constraints as needed.

We emphasize that validity is closely intertwined to sensitivity and reliability for AI as interviewer systems, in ways that map onto traditional concerns about construct validity. At the scale of automated interactions, reliability problems can directly undermine validity: the system may be personalizing to each *individual* respondent, yet it must standardize the interaction such that the same construct is being measured *across* all respondents. If the interaction varies in ways that change respondents' interpretation of the question, then respondents may no longer share a uniform understanding of what is being asked – at best introducing additional measurement error, and at worst yielding an incoherent construct (Conrad and Schober 2000, Kuha et al. 2018). To address this, in addition to pretesting and review, researchers would benefit from quantitative assessments of reliability, for example by measuring intra-interviewer correlations where each distinct model configuration is treated as an interviewer (see West et al. 2016).

The evaluation of AI as an interviewer differs in an important way from evaluating AI as a Colleague involved in questionnaire design. When LLMs are used to write, edit, or translate questions, researchers can review and revise the resulting artifacts before deployment. In contrast, when an LLM conducts interviews, it generates content dynamically in real time, making it impractical or impossible for researchers to review the full set of interactions before or even after fielding. As a result, evaluation must focus less on the static outputs and more on the process of deployment itself. This creates a stronger need for ongoing monitoring, testing, and governance to ensure that the LLM behaves consistently, remains within predefined bounds, and does not introduce unsafe, inappropriate, or scientifically invalid interactions with respondents.

4.3.3 AI as a Respondent

As detailed in §3.1.2, researchers may opt to use AI to generate simulated human response in various ways. This may take three forms. First, simulated responses may be created to test a survey instrument for skip logic issues, project frequency distributions, or other similar pre–data collection tasks. In this scenario, AI is similar to a colleague in that a research assistant would typically test a survey instrument for skip logic, and a subject matter expert may set some expectations for projected distributions of responses. Second, researchers may use generative AI to simulate responses for item-missing values, conceptually similar to imputation. Finally,

some researchers are replacing humans with synthetic respondents. In this case, the LLM or researcher chooses a persona and answers the questions as it believes that persona would, producing either individual responses or sample distributions of synthetic survey data.

In the scenario where AI is used for testing, considerations for the four criteria are largely the same as they would be in §4.3.1. We focus on the latter two use cases of imputation and synthetic responses since they are primarily measurement tasks, which have their own particular considerations across the four criteria.

Validity. Methods for checking construct validity vary considerably across these different use cases. When AI is used as a respondent for instrument testing, validity checks should focus on whether the system behaves as a real respondent would. At minimum, this means correctly following question intent, response constraints, and survey logic, including skip patterns. Researchers may opt to check for internal consistency and logic among responses. For example, in imputing prescription drug use, the LLM should not check "yes" for birth control for a 65-year-old woman.

When used for *imputation or prediction*, validity hinges on whether outputs are feasible and internally coherent rather than implausible, nonexistent, or accompanied by extraneous commentary. External sources may also be used to check basic distributions and correlations. Suppose that (at the time we wrote this) a survey of US adults did not ask political party identification on a survey, and an LLM was used to impute responses for all individuals. The overall distribution should be split relatively equally between Democrat and Republican and skew Democrat for women, younger individuals, and minorities, as is consistent with other contemporary survey data of the US adult population. Traditional quality metrics used for human respondents (e.g., tests for straightlining, primary/recency effect) are also applicable here to check whether the LLM is biased.

In *synthetic sample* applications, validity is best evaluated by whether outputs resemble realistic human responses from the target population, not merely surface-level plausibility, stereotypes, or artifacts of the model's training data–what Argyle et al. (2023) refer to as "algorithmic fidelity." Practical diagnostics include checks for unrealistically high refusal or "don't know" rates, selection of response options that do not exist, implausible or contradictory combinations across items, and irrelevant outputs such as explanations or meta-commentary (Lyman et al. 2025). There may be a need for additional checks for leakage or memorization (e.g., unusually specific strings reproduced verbatim from training sources) may be warranted.

When human respondents are replaced with synthetic responses, it's ideal that the data collection includes at least some human respondents, or a mixed-subjects design (Krsteski et al. 2025, Broska et al. 2025). This will allow comparisons of distributions and correlations between the simulated and human respondents to further evaluate the validity of the AI model as a more informed test of algorithmic fidelity. In practice, researchers may adopt a *validate-then-simulate* workflow (establish acceptable alignment before scaling synthetic generation) or, when constraints require, *simulate-then-validate* (Hullman et al. 2025). Parallel benchmarking may be

especially important for rapidly evolving or newly emerging topics, where prior human benchmarks may be outdated or unavailable.

Performance. In the domain of AI-as-a-respondent, performance concerns how accurately an AI system produces outputs from the target population, conditional on the task being well specified. In some survey applications–most notably item nonresponse imputation or full synthetic generation–individual-level accuracy is inherently unknowable because true values are unobserved. In these settings, performance assessment necessarily shifts from pointwise correctness to indirect diagnostics such as comparison with other statistical imputational models or distributional comparisons with external benchmarks.

Depending on the researcher's goals, certain statistical properties or machine learning metrics may require more focus than others. For instance, if the researcher intends to share aggregated cross-tabs or point estimates from a synthetic sample, "distributional fidelity"–as measured by item-level correlations with parallel human-based estimates or prior data–may be most important. However, if the researcher intends to disseminate individual-level data, then it is important to demonstrate respondent-level accuracy for each item; depending on the question and its particular format and properties, this may demand a shift toward metrics such as precision, recall, or mean squared error (MSE). Here, incorporating a mixed-subjects design is particularly valuable: it enables "rectification" of synthetic estimates using human responses to improve performance on metrics such as MSE.

Sensitivity. Assessing sensitivity is especially relevant when AI is used as a respondent. Different models will certainly have different methods for constructing responses, all of which are black boxes. The variance of methods may yield significantly different results, especially in tasks that do not have clear truths. Additionally, the same model may produce different results based on changes in user behavior. For instance, some researchers may feed in questions individually so future answers are not conditioned on prior answers (likely resulting in more inconsistencies for a given respondent); some may provide demographic profiles of human respondents; some may allow volunteered responses (e.g., don't know or refused). Researchers will want to understand how different processes alter the distributions and correlations of their variables to understand the effect they may have on the conclusions of their analysis.

Reliability. Because LLMs are quickly evolving in how they generate responses, reliability is likely to change over time. Within a year, LLMs went from consistently producing the median of a response distribution for a given set of demographics to sampling from the distribution (Sivaprasad et al. 2023). The former produced no variance while the latter does. Moreover, the LLM may learn from the data being input, learning from earlier profiles/responses. While this may be beneficial to improve accuracy of tasks such as imputing item-missing values, it may violate assumptions of several analytic methods (e.g., assumption of independence, assumption of uncorrelated errors) which could skew conclusions drawn from analysis of these data. Researchers may use test prompts/data to replicate the LLM tasks at different points in the same data collection and across data collections to ensure consistent results. With that said,

some changes are expected over time due to true change in the population. Researchers will need to use external comparison points to determine whether changes across time are true change or model evolution. Additionally, estimation frameworks such as prediction-powered inference (PPI) enable more appropriate uncertainty quantification for mixed samples (Angelopoulos et al. 2023), and human-in-the-loop validation frameworks can enable catching errors and drifts from previous observations (Vishwakarma et al. 2024).

4.3.4 AI as an Analyst

After data are collected, substantial analytic work is typically required before conclusions may be drawn. As described in §3.2, AI systems, particularly LLMs, may be used to perform or assist with these tasks.

In qualitative research, this may include transcription, coding, and thematic synthesis. In quantitative research, it often involves data cleaning, recoding, and the interpretation and communication of statistical results. In both cases, researchers make a series of interpretive and procedural decisions that shape analytic inputs and outputs but are not themselves acts of statistical estimation. Unlike previous example-use cases, data cleaning and analysis often follow hard rules (e.g., if someone says Cuban, back code as “Hispanic;” a 5th grade teacher belongs in the “education” industry). Similar to human coders, LLMs may be used to code open-ended responses into a close-ended set of options. Finally, AI may be used to assist or carry out analytical duties such as data cleaning, documentation, or data visualization; though these are tasks that typically fall to an analyst, considerations for this specific usage align closely with the AI-as-a-colleague.

Validity. For thematic coding and labeling, task validity depends heavily on how the LLM is instructed to interpret the codeframe and apply it consistently. Many instructional strategies exist. For instance, *zero-shot prompting* explicitly defines the codebook (categories, decision rules, and exclusions) and provides relevant project context (e.g., target population, survey mode, and scope definition of each code). In some settings, prompts may also incorporate hard constraints (e.g., “only output one category label from this list”) or rules that operationalize the codebook. *Few-shot prompting* augments this approach by providing concrete examples, ideally covering typical cases, edge cases, and common confusions between categories (Brown et al. 2020). These approaches can be further strengthened by adding guardrails such as expected base rates (when known), explicit “do not infer beyond text” instructions, and a low-confidence or “needs human review” option for ambiguous cases. Across prompting strategies, validity checks should emphasize whether outputs are interpretable, comprehensive, and aligned with the intended construct (e.g., avoiding category drift, definitional errors, or systematic overuse of catch-all codes).

Beyond back-coding, AI is increasingly used for analytic tasks (including “coding agents” that execute multi-step pipelines). For AI-as-analyst workflows, validity depends not only on the user’s instructions, but also on the *configuration of the toolchain* and the information environment the system can access. Whether an agent correctly performs cleaning,

transformation, or visualization tasks will hinge on which tools it can call (and how), what datasets and folders it can read or write, and what supporting documentation it is given (e.g., data dictionaries, READMEs, analysis plans). Validity checks in these contexts should therefore assess both (a) whether the outputs match the intended analytic task, and (b) whether the agent used appropriate inputs, assumptions, and transformations given the available documentation and constraints.

Performance. For back-coding and labeling tasks, performance is often evaluated using agreement-based metrics (e.g., accuracy, precision/recall, F1 score, or area under the curve) when a human-coded benchmark is available. Recent studies suggest that LLMs can achieve promising performance across multiple dimensions of back-coding and classification, particularly for well-defined codeframes and high-frequency categories (e.g., Mellon et al. 2024, von der Heyde et al. 2025). Still, strong aggregate performance can mask uneven accuracy across subgroups, rare categories, or edge cases, underscoring the importance of disaggregated evaluation where feasible. It is also important to recognize that many coding tasks are inherently subjective: open-ended responses may be ambiguous, span multiple categories, or lack sufficient detail to support a single “correct” classification. As a result, performance thresholds should be interpreted in light of task difficulty, coder agreement levels, and the substantive consequences of different types of error.

For broader analytic tasks or multistage workflows performed by coding agents, performance evaluation is more challenging due to the multidimensional and open-ended nature of outputs. In such cases, evaluation often relies on comparisons to replicated final outputs produced by established pipelines, targeted unit tests for specific steps, or structured human review. These assessments may themselves involve judgment and are often impractical to conduct systematically, particularly for one-off or exploratory uses. As a result, researchers may need to rely in part on upstream performance evaluations (e.g., agentic benchmarks or prior validation studies), while recognizing that performance in a specific applied context may differ from benchmark settings. Across all use cases, the most appropriate performance metrics will depend on the task, available ground truth, and tolerance for different types of error.

Sensitivity. Sensitivity concerns the extent to which AI outputs change in response to reasonable variations in inputs, rules, prompts, or system configuration. For labeling and back-coding tasks, LLMs may produce different results depending on how coding instructions are specified. For example, results can vary based on whether uncertain cases are forced into a category or deferred for human review, whether multiple labels are allowed, and how ties and borderline cases are resolved (von der Heyde 2025). Sensitivity may also arise from seemingly minor prompt variations (e.g., wording, ordering of instructions, or inclusion of examples), changes in model version, or differences in decoding or reasoning settings. These differences do not necessarily reflect errors, but rather alternative operationalizations of the same underlying construct. In some contexts, it may reveal genuine ambiguity in the data or codeframe.

In agentic or multi-stage workflows, sensitivity can compound across steps, as early-stage variations propagate through downstream transformations or analyses. This is particularly salient when AI systems are used to mine the same open-ended responses for multiple analytic purposes, which can amplify measurement and processing error if assumptions are not carefully aligned across tasks. Assessing sensitivity typically involves stress-testing the system under plausible alternative specifications, such as running multiple prompts, toggling decision rules, or comparing outputs across models or configurations, and examining whether substantive conclusions remain stable.

Reliability. Reliability refers to the consistency of AI outputs over time, across repeated runs, or under nominally identical conditions. As LLMs' language capabilities continue to evolve, their coding and analytic performance may improve, but this evolution also introduces challenges for replicability. Unlike traditional statistical models, where fixed code applied to the same data yields identical results, repeated back-coding with different model versions, or even the same model under updated system parameters, may produce different classifications and, in turn, different substantive findings (Barrie et al. 2024).

Where possible, researchers should attempt to bake in explicit assumptions into their model inputs to reduce unwanted variation, including fixed prompts, documented decision rules, versioned codeframes, and controlled model settings. Using open-source or locally hosted models with carefully monitored updates can further support reproducibility, though such approaches may trade off against rapid performance gains available in proprietary systems. At the same time, improvements in model accuracy may legitimately manifest as shifts in estimated distributions or category assignments, in which case newer models may be preferable despite reduced comparability to earlier results.

Given these tensions, reliability should be treated as an ongoing property of the AI-as-an-analyst usage model, rather than a one-time certification. Periodic re-evaluation, especially when frontier models or toolings change, is essential for understanding whether observed differences reflect meaningful improvements, shifting biases, or instability in the measurement process. For coding tasks in particular, the current empirical literature on long-term reliability remains limited, reinforcing the need for continued documentation, benchmarking, and cautious interpretation as agentic tools mature.

4.4 Other Evaluation Considerations

The four evaluation criteria described above (validity, performance, sensitivity, and reliability) provide a general framework for empirically assessing the quality of AI usage in survey research. However, depending on the specific task or application, additional criteria may also be relevant. These criteria may be particularly important when evaluating AI systems embedded within survey workflows, respondent interactions, analytical pipelines, or reporting tools. Table 3 enumerates some of these additional criteria and examples where their usage may be appropriate.

Table 3. Additional criteria for task-specific evaluation.

Criterion	Definition	Example
Simplicity	The degree to which outputs or solutions avoid unnecessary complexity while still achieving the intended objective.	When GenAI is used to generate draft survey questions, parsimonious wording may be preferable to ensure clarity and respondent comprehension.
Generalizability	The extent to which a GenAI system's outputs or performance remain effective across different populations, datasets, or survey contexts.	A model used to categorize open-ended responses should ideally perform evenly across demographic groups or survey waves.
Security	The ability of the system to protect sensitive information and resist unauthorized access or manipulation.	GenAI systems used to process respondent data should not expose personally identifiable information through prompts, logs, or outputs.
Interpretability	The degree to which researchers can understand how or why a system produced a particular output.	When GenAI is used for modeling or estimation, researchers may prefer approaches that allow inspection of the factors driving predictions.
Trustworthiness	The extent to which the system behaves in a predictable, responsible, and accountable manner (as deemed by key stakeholders).	Researchers may prefer a GenAI model for the role of AI as an interviewer that has demonstrated a level of public trust according to user research studies.
Automaticity	The extent to which a task can be performed with minimal manual intervention while maintaining acceptable quality.	Researchers may prefer GenAI systems using models that are able to achieve an acceptable level of open-ended coding accuracy with minimal human review.
Persuasiveness	The degree to which outputs influence beliefs, interpretations, or decisions.	AI-generated summaries used in research briefs should avoid overstating certainty or presenting speculative interpretations as established findings.
Safety	The extent to which GenAI system outputs avoid causing harm, including the generation of misleading, biased, or inappropriate content.	GenAI used in respondent-facing interactions should avoid producing offensive or misleading prompts.
Realism	The degree to which generated outputs plausibly resemble human-generated responses or behaviors when that is required for the task.	When GenAI is used to simulate respondents for testing survey instruments, responses should resemble plausible human answers.
Efficiency	The extent to which GenAI improves the speed or cost-effectiveness of a workflow without compromising quality.	Using GenAI to draft initial analytic summaries may reduce the time required for researchers to prepare reports.

More broadly, some of the criteria listed in Table 3 may not represent wholly distinct dimensions of evaluation so much as task-specific elaborations of the four core criteria introduced above. In some cases, they may simply express what validity requires for a particular task: for example, *realism* may be part of establishing validity when the task is to generate plausible synthetic responses, while *interpretability* may be central to validity when the task involves supporting human judgment or explanation. In other cases, these considerations may serve to operationalize performance, identifying what "better" or "worse" output means in a given application. *Efficiency*, *persuasiveness*, or *simplicity* may matter not as universal desiderata, but as application-specific ways of assessing whether the system is performing the intended task well. Still other considerations may overlap with reliability or sensitivity. *Safety*, for instance, may partly concern reliability if harmful failures occur inconsistently or unpredictably, and it may also concern sensitivity if outputs deteriorate under plausible changes in prompts, contexts, or user populations. For these reasons, researchers should not treat the criteria in Table 3 as a rival framework to the four core criteria above, but rather as supplementary, task-dependent lenses that may help specify what those broader criteria mean in practice for particular survey applications.

We encourage researchers to draw on criteria from other established evaluation frameworks when those criteria help more precisely operationalize the relevant dimensions that matter for a given task.[8]

Finally, we emphasize that a key consideration that cuts across all evaluation efforts is transparency. While transparency is not strictly a criterion for assessing output quality in the same sense as the categories above, it plays a foundational role in enabling meaningful evaluation. Without sufficient transparency about the models used, prompts provided, tools employed, and human oversight applied, it becomes difficult for others to assess whether reported claims of validity, performance, sensitivity, or reliability are credible. For this reason, transparency may be viewed as a "fifth category" of evaluation, one that supports all others. We therefore turn next to recommendations for transparency and reporting.

5. Recommendations for Transparency and Reporting

AAPOR has long held that transparency and disclosure are foundational to credible survey research. Researchers are expected to document how their data were collected, how samples were constructed, and how results were weighted, not merely to satisfy reviewers, but because these disclosures are what allow consumers of survey data to assess quality, identify limitations, and place appropriate confidence in findings. As AI becomes embedded in the survey lifecycle,

[8] See, for instance, Wilkinson et al. (2016), Mitchell et al (2019), Amaya et al (2020), Wagner et al. (2021), NIST (2024), Rabanser et al. (2026), Rao et al. (2026).

that same commitment to transparency must extend to how AI was used, where it operated autonomously, and how human judgment shaped its outputs.

The disclosure framework proposed by this task force is built on three interconnected ideals: **replication**, **reproducibility**, and **understanding**. Replication—achieving the same result under identical conditions—is almost always unattainable in surveys given random variability and temporal variation and is even less likely when AI is involved. Reproducibility—obtaining similar results under similar conditions—remains a critical ideal for some survey data users, especially academic researchers. For many survey consumers, however, the primary transparency goal is understanding: providing enough detail to recognize potential bias and limitations.

At its core, the disclosure framework serves three purposes: **(1) transparency**, by clearly indicating where AI was in the loop and how human judgment shaped the process disclosure supports both reproducibility and interpretability; **(2) creating data for science**, by documenting practices disclosure enables the research community to learn how and where AI is being used and to better anticipate the future of survey research; and **(3) guiding researchers during their work**, knowing that disclosure is required encourages researchers to treat it not as an afterthought, but as a tool that informs decisions throughout the research process. Section §3 complements this discussion with concrete AI use cases in survey research, and section §4 provides corresponding evaluation techniques.

5.1 Disclosure Checklist for the Use of AI in Surveys

The disclosure checklist introduced in this section is more than an evaluation instrument, it is a roadmap for responsible and transparent survey research in an era of AI. It is organized into two tiers:

- **Required disclosures** should be included in any reporting or methodological summary and be presented in a way that is clearly disclosed and easily accessible to readers. The required disclosures represent the minimum information necessary for consumers to understand potential bias and limitations. As of publication of this report (May 2026), the required AI disclosures have been submitted as proposed revisions to the AAPOR code and are pending a membership vote.
- **Enhanced disclosures** are necessary for reproducibility and are strongly encouraged to support the goals of transparency, creating data for science and informing research decisions. We urge journals to adopt them as a condition of publication.

We recognize that no disclosure framework can anticipate every future development of AI tools, workflows, or applications. The goal of the taskforce was not to create a permanent form, but one that can be flexible and updated to evolve with technology while specific enough to be actionable today. This checklist and disclosure requirements will need to adapt in the future. Transparency requirements cannot be static when the technology itself is not.

Although AI often performs tasks previously carried out by humans, it introduces sources of error that are not adequately captured by pre-AI reporting standards. For example, stating that "interviews were conducted by trained call-center staff" is sufficient because consumers share well-understood expectations about human interviewer training and error. By contrast, simply stating that "interviews were conducted by an AI system" tells consumers very little. AI systems exhibit a much wider range of potential behaviors, failure modes, and biases, many of which are opaque or unfamiliar. These disclosures exist not to treat AI as categorically different from prior technologies, but to ensure that when AI replaces or augments human labor, reviewers and consumers have enough information to meaningfully assess what that means for data quality and validity.

Table 4 (Required disclosures) and Table 5 (Enhanced disclosures) lay out each required and enhanced disclosure item, what to report, and why it matters. For each item in the checklist, researchers should provide an answer; if an item cannot be answered, researchers should explain why (for example, the information is unknown or disclosure would jeopardize respondent confidentiality). To ease the burden of compliance, we are developing an online tool that will streamline completion of the checklist. Assuming researchers have documented their AI use throughout the project, the tool is designed to take just a few minutes and can also support the preparation of broader methodological descriptions.

Journals, IRBs, and relevant laws and regulations (e.g., CCPA/GDPR) may have their own reporting requirements beyond those listed here. The disclosure list is intended to complement, rather than override, those requirements.

5.1.1 Required disclosures

The required disclosures establish the minimum standard for documenting how AI was used in a study. This information is necessary for readers, reviewers, sponsors and clients to understand how the research was conducted and where bias may have been introduced. Examples of methodology statements that comply with the minimum required disclosures can be found in section 5.2. Readers will note that these requirements are not expected to add undue burden to the reporting process and a few sentences are adequate for meeting the reporting requirement.

If AI was used as an **Interviewer**, **Respondent** or **Analyst,** the researcher must disclose the task performed by the AI, whether there was human oversight or validation of the AI, and the number of human respondents (if any). It is worth noting what these disclosures do **not** require. The most common tasks of using AI as a Colleague (for example, in helping to ideate the survey instrument) or AI as a Briefer (for example, in augmenting the writing of the report) are not generally required disclosure as part of our recommended minimum standards, but the researcher may choose to do so (or it may be requested by the consumer). The artifact of their work (for example, the survey questions or the report) is easily investigated and evaluated by a consumer directly, making it less necessary to understand the backstory of their creation.

Human oversight is a critical safeguard against AI-driven errors, including misinterpretation of respondent input, inappropriate interviewer behavior, biased coding, or analytic mistakes that

may not be apparent from final outputs alone. Disclosure of when and how oversight occurred enables data users to assess whether such errors were likely to have been detected and corrected. More detailed descriptions of validation procedures, (e.g., independent review, structured audits, or statistical validation) signal greater rigor in error control and governance. In the absence of this information, consumers cannot distinguish between AI outputs that were systematically verified and those that were effectively trusted without review.

Accordingly, these guidelines require disclosure of both whether AI outputs were validated and how that validation was performed, but does not prescribe specific validation methods. Appropriate oversight varies substantially by task, research context, and risk profile, and will continue to evolve alongside advances in models, infrastructure, and empirical evidence on validity. Our goal is therefore not to standardize validation practices, but to ensure transparency sufficient for informed assessment, replication, and appropriate use (Suh, Smith, and Chit 2025).

Reporting the number of human respondents clarifies the balance between human input and automation, which is essential for evaluating data quality, error mitigation, and interpretability. Human respondents remain a relatively stable and meaningful unit of analysis. By contrast, reporting the number of AI instances is neither required nor recommended: this quantity is often ambiguous, non-independent, and highly sensitive to implementation details (e.g., batching, retries, or agent architecture), and therefore provides little insight into bias, validity, or data quality risk.

Requiring disclosure of human involvement, while avoiding misleading AI counts, supports transparency without imposing reporting burdens that are unlikely to improve scientific inference (Argyle et al. 2023). We acknowledge that this standard assumes a clear boundary between human and AI-generated responses. Future advances such as digital twins or hybrid human–AI constructs may increasingly blur this boundary, warranting revisitation of these requirements as the technology evolves.

How AI-generated data are described shapes whether readers can accurately assess what a study actually measured. Generated responses or data that are generated, inferred or modeled through artificial intelligence (e.g., silicon responses, digital twins, synthetic responses) are not research “participants.” Cases created in this manner and included in a purported study of public opinion must be identified as having been created through artificial intelligence. “Poll,” “polling,” “survey,” and “surveying” and other similar terms imply that the primary source of data are from human respondents. These terms should not be used to describe data created through artificial intelligence. Section 5.2 has example language.

Table 4. Required disclosures

Disclosure item	What to report	Why it matters
Tasks performed by AI	Describe which roles AI played, if it played any of the following roles*:	The meaning of validity, performance, sensitivity, and

	• **Interviewer** - Asking questions (e.g., a voice-enabled AI chatbot administered the survey, adapting follow-up questions based on prior responses). • **Respondent** - Simulating the target population. • **Analyst -** Cleaning, labeling, or modeling data. • **Other -** Any other role the researcher chooses to disclose.	reliability differs fundamentally by task. Without knowing *what* AI did, consumers cannot assess where errors may have entered or how to interpret results.
Description of AI's role	Briefly describe in plain language what the AI did. *Example: "A voice-enabled AI chatbot administered the survey, adapting follow-up questions based on prior responses, while maintaining neutrality and standardized delivery."*	A task label alone (e.g., "AI was used as an interviewer") does not convey enough detail. A brief description lets the research consumer gauge the scope of AI involvement.
Human oversight or validation	Researchers should disclose whether there was human oversight or validation of the AI. • **Validation/oversight tasks** - For which task(s) was the AI procedure or output reviewed or validated by researchers? (e.g., AI as an Interviewer, respondent or analyst) • **Validation details** - Researchers should report how oversight was conducted (e.g., manual review by researcher, statistical check, cross-validation with human coded data, linking to existing or previously conducted validation or technical documentation).	Human oversight is a critical safeguard against AI-driven errors. Without knowing whether and where humans reviewed outputs, consumers cannot distinguish between results that were systematically verified and those that were effectively trusted without review. The rigor of oversight can vary enormously. Describing the method allows consumers and reviewers to evaluate whether the level of validation was appropriate to the task and its risk.
Human respondents	Researchers should disclose: • **Number of human respondents -** Report the total number of humans who responded to the survey. • **Synthetic data -** Synthetic data must be clearly described as such. Generated responses or data that are generated, inferred or modeled through artificial intelligence (e.g., silicon samples, digital twins, synthetic samples) are not research "participants." Cases created in this manner and included in a purported study of public opinion must be identified as having been created through artificial intelligence.	Reporting human involvement clarifies the balance between human input and automation. This is also essential for evaluating the meaningfulness of any reported sample sizes or confidence intervals. Synthetic responses do not have the same properties as random samples. Their biases are systematic and may be invisible to standard quality checks. Consumers need to know when and how synthetic data were used in order to interpret results appropriately.

*Note: Use of AI as a Colleague or as a Briefer does not generally require disclosure, since the resulting artifact (the survey or report) can be directly reviewed by consumers.

5. .1.2 Enhanced disclosures

Where required disclosures establish the minimum standard for transparency, enhanced disclosures provide the additional detail necessary for reproducibility and deeper scrutiny. We urge journals to adopt them as a condition of publication. As with required disclosures, tasks where AI is the **interviewer**, **respondent**, or **analyst** should be disclosed, while tasks where the AI is a colleague or briefer do not typically require disclosure.

These items should be considered throughout the research process, and not assembled after the fact, both because it helps ensure a healthy research workflow and some of these questions are hard to answer retrospectively. The enhanced disclosures are organized into two groups: information about how the AI system was accessed and configured and information about the specific model, prompts, and instructions used. All researchers should be able to answer questions related to access and configuration, while only a subset (those using first-party tools or highly transparent third-party tools) are expected to provide detailed model-level information. Together, these disclosures provide sufficient context for informed evaluation of AI-mediated survey research as models, tooling, and norms continue to evolve. They are not intended to prescribe a single correct approach, but to ensure that the choices researchers made and the constraints they worked under are visible to those who use and evaluate their work.

We recognize that researchers using third-party proprietary tools or AI systems embedded within survey platforms may not have access to many of these details. Model versions, configurations, and system-level instructions are often not disclosed by vendors. In these cases, researchers should provide as much information as they can, note explicitly what is unknown and why, and link to any documentation made available by the third-party provider. We strongly encourage platform providers and vendors to improve their transparency (Wan et al. 2025), keep updated documentation about their systems, and to make this documentation readily available to their research clients and users.

Table 5: Enhanced Disclosures

Disclosure Item	What to Report	Why it Matters
Access and Infrastructure		
Method of access	Disclose how the AI model was accessed: ● Direct access (e.g., API) ● First-party platform or tool (e.g. a provider-hosted chatbot or interface) OR ● Embedded within a third-party platform or tool *(see next item)*	Access methods shape researcher control, transparency and reproducibility and may introduce distinct sources of error. For example, an interviewer bot embedded within

		a survey platform may generate interaction- or interface-driven biases that differ meaningfully from those arising in stateless API calls. Understanding the tooling context helps reviewers identify risks related to platform constraints or limited configurability.
Third-party instrument or interface *(third-party only)*	If within a third-party platform or tool: ● **Instrument or interface:** report where and how the AI was embedded or interacted with (e.g., Qualtrics integration, custom dashboard, interviewer bot). a. **Disclosure possible:** Indicate whether the provider discloses relevant model details (yes / no / don't know). If the answer is "no" or "don't know," researchers may skip questions about "Model Details and Core Prompts and Instructions"	Third-party interfaces may modify model behavior in ways that are not visible to researchers. Identifying the specific interface helps readers assess what constraints or modifications may have been applied. Researchers often cannot access model details when using embedded third-party tools. Flagging this explicitly, rather than leaving fields blank, signals to consumers that the limitation exists and was acknowledged.
Dates of access/use	Report when the AI system was used for the task (e.g., November 18, 2025).	Reporting the date(s) of access or use situates the study within the state of available models, tooling, and norms at the time of data collection, an essential context given the rapid pace of change in AI systems.
Memory/ statefulness	Disclose whether the system was stateful (whether prior interactions were retained and allowed to influence future outputs or stateless during interaction).	Stateful systems may introduce dependence across interactions, while stateless systems may reduce such risks at the cost of coherence or task performance (e.g., does an AI interviewer evolve their questions based on earlier answers).
Full interview archive *(AI as an Interviewer only):*	Researchers should indicate whether interview archives or transcripts can be released.	When AI is used as an interviewer, the full archive of interactions is the primary

		mechanism for after-the-fact assessment of variability, comparability across respondents, and measurement consistency.
Known biases	Document any known biases associated with the model, platform, or access method that could plausibly affect results.	Explicit acknowledgment of known biases, documented failure modes, and human overrides further supports transparency and evaluability.
Justification for model and access choice	Brief explanation of why the specific model and access method were chosen. Relevant considerations may include performance, transparency, reproducibility, ethical considerations, cost, ease of use, or lack of feasible alternatives.	Choices about which model to use are not neutral. Different models carry different biases, capabilities, and governance structures. Disclosure allows consumers to evaluate whether the choice was appropriate and whether alternatives might have produced different results.
Model details and core prompts and instructions (*only asked if access was direct, first-party tool, or model details are disclosed in third-party tool*)		
Details about the model name, version and type	Report the specific AI system used, including the model name and version (e.g., GPT-4.0, GPT-5). Include: ● **Model type:** indicate whether the model is open-source (publicly available) or proprietary (owned or controlled by a company). ● **Source URL:** provide a link to official model documentation, where available.	Even minor version updates can meaningfully alter model outputs, biases, and alignment. Without a specific version, results cannot be reproduced or compared across studies, and readers have no basis for evaluating what the system was capable of or likely to do. Proprietary models may limit visibility into training data or updates, while open-source models may allow deeper scrutiny; disclosing model type helps situate these tradeoffs (Mitchell et al. 2019; Bender et al. 2021). Direct links reduce ambiguity about which system was used and give readers a path to further information about capabilities, limitations, and known issues.

Fine-tuning status	Disclose whether the model was fine-tuned for survey-related tasks (yes/no). ● **Fine-tuning details:** If the model was fine-tuned, researchers should describe the data used and its source(s). Where possible, links to the dataset or citations to the data source should be provided.	A fine-tuned model may behave quite differently from its base version, particularly in how it classifies responses or handles sensitive topics. Without disclosure, consumers cannot assess whether the model's behavior reflects general training or domain-specific adaptation.
RAG usage	Disclose whether and how RAG was used to ground model outputs, including the types of documents or resources retrieved (e.g., a catalog of approved survey questions used by an interviewer chatbot).	RAG systems ground outputs in external documents, which can improve factual accuracy but also introduce biases from the retrieved content. Disclosing RAG usage allows consumers to evaluate the quality and representativeness of the retrieval corpus.
Custom configurations	Any non-default settings that could plausibly affect outputs, such as temperature, maximum token length, random seed, or number of runs. *Example: temperature = 0.7; max tokens = 2,000; seed = 42; runs = 3.*	Small changes in configuration can produce meaningfully different outputs. Reporting these settings is essential for reproducibility and helps readers understand the degree of determinism or randomness in the outputs.
Prompts	Report the prompts used to guide the model. Exact prompts are preferred, but high-level, plausibly abstracted descriptions are acceptable when exact prompts cannot be shared.	Prompts and system-level instructions are a central determinant of AI behavior. Even subtle differences in wording, constraints, or emphasis can introduce systematic variation or bias in outputs. For example, instructions favoring brevity may suppress nuance in open-ended responses, while prompts encouraging inference may induce over-interpretation. Reporting prompts and configurations therefore enables reviewers to evaluate whether observed patterns reflect substantive phenomena or artifacts of model instruction. (Abraham, Arnal, and Marie 2025)

System-wide instructions	Disclose any global instructions or settings that governed the model's behavior across tasks.	System prompts establish the operating constraints for the entire interaction. They shape tone, boundaries, and priorities in ways that affect all outputs. Without knowing what system-level instructions were in place, readers cannot fully interpret individual outputs or task-level results.
Code	Report any scripts or code used to call or orchestrate the AI system, where feasible.	Code disclosure supports reproducibility.

5.2 Disclosure Examples/Vignettes

These vignettes are intended to help provide examples of how AI may be used in the research cycle and how researchers can meet the **required standards for disclosure**, including the use of AI. AAPOR is also developing an online checklist, which will allow researchers to easily enter the details of their study, including the use of AI. The responses will be recorded and can be used to write a methodological statement or create an appendix or list of enhanced disclosures.

Example 1: AI was used to conduct telephone interviews and for open-ended coding. (AI as an Interviewer AND AI as an Analyst)

Results for this poll are based on telephone interviews conducted Dec. 1–15, 2025, with a random sample of 1,000 adults, aged 18 and older, living in all 50 US states and the District of Columbia. Interviews were administered using a voice-enabled artificial intelligence (AI) interviewing system that delivered standardized survey questions by telephone. Human researchers monitored system performance and reviewed a subset of interviews to ensure adherence to question wording and neutrality.

For results based on the total sample of national adults, the margin of sampling error is ±4.0 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

The sample had a quota of 80% cellphone respondents and 20% landline respondents. Landline and cellular telephone numbers were selected using random-digit-dial methods.

Open-ended survey responses were initially coded using AI-assisted text classification. Human researchers reviewed, validated, and, where necessary, corrected AI-generated codes prior to analysis.

Question wording, AI-assisted interviewing and coding processes, and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls. For more

methodological details, including information about the use of AI, please refer to this supplementary documentation [hyperlinked].

Example 2: AI was used to generate synthetic cases. (AI as a Respondent)

These results are based on 500 synthetic responses that were generated on December 1, 2025 using an AI model. The responses generated were intended to be representative of the U.S adult population including age, gender, education level, race/ethnicity, party ID, and region. These procedures were conducted using [insert product name]'s AI-driver simulation methods, which applied advanced machine learning and large language models to recreate respondent profiles and survey responses based on real-world indicators such as recent news and social media content.

Researchers from [insert organization name] reviewed model specifications and output distributions for quality assurance prior to analysis of the data. For more information about the model, including prior validation work, please refer to the documentation that can be found here [insert link].

A margin of error cannot be produced from synthetic responses. The use of AI-based simulated responses, along with the question prompts used to generate the data, may introduce error or bias into the findings. Users of these data should interpret results with full awareness of the role and limitations of the AI methodologies employed.

Example 3: AI was used to conduct, transcribe and code qualitative interviews. (AI as an Interviewer)

This study is based on 500 in-depth qualitative interviews conducted between March 3 and April 12, 2026, with English speaking adult participants aged 18 and older residing in the United States. All interviews were conducted with human respondents recruited from [organization name]'s probability-based panel. Interviews were approximately 60 minutes long.

Interviews were administered by a conversational artificial intelligence (AI) interviewing agent using a secure, voice-based platform. All respondents consented to being interviewed and recorded by the AI system. The AI agent followed a semi-structured interview guide developed by the research team, which included core questions, suggested probes, and topic transitions. While the agent adhered to standardized wording for primary questions, it was permitted to ask contextually appropriate follow-up questions and probes in response to participant answers to encourage elaboration and clarification. The AI system was tested extensively prior to fielding, including pilot interviews with internal volunteers and iterative review of probe behavior, to ensure neutrality, consistency, and adherence to the interview guide.

All interviews were audio recorded and transcribed automatically by the AI system. Transcripts were then analyzed using AI-assisted qualitative coding, in which the system generated preliminary thematic codes and summaries based on patterns observed across responses. A

detailed codebook was developed collaboratively by the research team prior to analysis and used to guide the AI's thematic classification.

Human researchers conducted validation and oversight throughout the qualitative analysis process. Approximately 10% of completed interviews were randomly selected for full human review, including listening to audio recordings, verifying transcript accuracy, and assessing the appropriateness of AI-generated thematic codes. Where discrepancies or misclassifications were identified, researchers revised the codes and used those corrections to refine the coding framework for subsequent analysis.

As with all qualitative research, findings are subject to limitations related to question wording, respondent interpretation, and interviewer-respondent interaction. In addition, the use of an AI-based interviewing and coding system may introduce unique sources of error or bias, including variability in probing behavior and challenges in interpreting nuanced or emotionally complex responses. Results should therefore be interpreted as illustrative of key themes and perspectives rather than as statistically generalizable estimates.

Example 4: AI was used to conduct the web survey and to carry out fraud detection (AI as an Interviewer and as an Analyst)

This study is based on a web survey conducted March 18–29, 2026, with 1,000 adult respondents aged 18 and older residing in the United States. Participants were recruited from [insert name]'s opt-in online panel and were screened to meet study eligibility criteria. Quotas were applied to approximate national benchmarks for age, gender, race and ethnicity, education, and region.

Data were collected using the [insert platform name] AI-enabled web interviewing system. Respondents completed a structured, self-administered web survey consisting primarily of closed-ended items. In addition to the standardized questions, the system was designed to ask optional, context-dependent qualitative follow-up questions after selected items. These AI-generated probes were intended to elicit brief explanations or clarifications of respondents' answers (for example, asking respondents to explain the reason for a rating or choice). Probing behavior was constrained by predefined rules to ensure relevance and neutrality. More information about this platform can be found here [insert link].

The survey instrument and AI probing logic were tested prior to fielding through multiple pilot runs, including internal testing and a soft launch with panel members, to evaluate question flow, probe frequency, and respondent burden. Human researchers reviewed probe content to confirm that follow-up questions were correctly executed and did not introduce different question contexts across interviews.

Following data collection, responses were evaluated using a separate AI-based fraud-detection and data-quality system designed for opt-in panel research. This system flagged potentially problematic cases based on multiple indicators, including outlier completion times, inconsistent or contradictory responses across related questions, repetitive or nonsensical open-ended text,

duplicate digital fingerprints across submissions, and response patterns commonly associated with automated bots or professional survey fraud. Flagged cases were reviewed by human researchers, who determined whether to retain or remove responses based on established data-quality criteria.

After quality review, the final analytic dataset included only respondents who passed both automated and human validation checks. Quantitative results are reported using standard descriptive and inferential statistics. Qualitative follow-up responses were summarized using AI-assisted text analysis, with human review of themes to ensure interpretive accuracy.

As with all opt-in panel surveys, results are subject to limitations related to nonprobability sampling and self-selection. In addition, the use of AI-enabled probing and automated fraud detection introduces potential sources of error or bias, including variation in follow-up questioning and imperfect identification of fraudulent responses.

5.3 Infrastructure and Audience Transparency and Reporting

For many researchers, decisions about AI tooling, models, and services are made indirectly through the survey infrastructure and audience providers they rely on, rather than through direct model selection or configuration. As AI becomes increasingly embedded within survey platforms and respondent marketplaces, critical design and deployment choices such as where AI is used, which models are involved, and how they interact with respondents or data may be opaque to researchers themselves. Without cooperation from these providers, researchers may lack access to information that is essential for responsible disclosure, evaluation, and interpretation.

Our concern extends beyond ex-post reporting for purposes of review or compliance. Transparency is also a prerequisite for learning. Researchers need visibility into infrastructure-level AI use to evaluate what works in their specific contexts, to diagnose unexpected behavior or failure modes, and to build cumulative knowledge about how AI affects survey measurement, data quality, and inference. When AI use is hidden or poorly documented at the platform or audience level, these learning processes break down, and errors or biases may persist undetected across studies.

In this subsection, we outline the roles of survey infrastructure providers (i.e., organizations that supply tools for survey design, deployment, and data processing) and audience providers (i.e., organizations that sell or broker access to respondent pools) in enabling a transparent and trustworthy survey ecosystem. While this framework does not impose requirements on these vendors, it highlights how provider-level transparency can substantially reduce disclosure burden for researchers, improve reproducibility, and support more informed evaluation by data users. We encourage researchers to be attentive to these considerations when selecting platforms and samples, and to weigh the scientific and ethical costs of opacity alongside more familiar tradeoffs such as cost, convenience, and scale.

Survey infrastructure providers play a central role in enabling, or constraining, transparency. As AI becomes increasingly embedded in survey platforms (e.g., in question generation, response validation, data quality assurance, or text summarization) platforms should clearly disclose the AI systems they deploy. While some uses of AI may be readily apparent to researchers, others may not be visible at all. Accordingly, infrastructure providers should document where and how AI is integrated into the survey workflow. For instance, if AI is used to flag low-quality responses, suggest edits to survey questions, or summarize open-ended responses, researchers should be informed whether these processes occur in real time, post hoc, or invisibly in the background.

At a minimum, such disclosures should include the model name and version, model type (e.g., open or proprietary), access method (e.g., API-based or embedded service), and fine-tuning status: particularly whether models are trained on user-provided or respondent data. This information is essential for reproducibility and auditability, and for understanding how model updates or drift during the data-collection period may affect research findings. To be practically useful, these details should be readily accessible to researchers, ideally through automated metadata exports or audit logs, enabling researchers to meet their own transparency obligations without needing to reverse-engineer the tools they use.

For audience providers, transparency regarding participant sourcing and validation is equally critical. Researchers should have access to aggregate indicators of participant activity, such as the number of surveys completed within a given time period, average time spent per survey, and dropout or attrition rates. These indicators help researchers assess risks associated with panel fatigue, satisficing behavior, or overexposure to particular types of studies, all of which may introduce bias. Importantly, they also help researchers identify AI-related risks, ranging from ecosystem-embedded autonomous agentic survey completion or the use of generative AI to produce or augment responses to open-ended questions.

Participant verification practices should also be transparently documented. Audience providers should disclose the methods they use to ensure that participants are real humans rather than bots or AI-generated personas (including so-called “digital clones”), and that participants reside and/or work in the locations they claim. Such methods may include IP-based checks, device fingerprinting, behavioral validation, or third-party identity verification. As synthetic responses become increasingly sophisticated, researchers should not be left to infer the reliability of their samples without clear information about participant validation practices.

To support researchers in meeting transparency standards, both infrastructure and audience providers should offer standardized reporting mechanisms or exportable metadata that capture relevant information about AI usage and participant validation. These reports should be designed to integrate directly with academic, regulatory, or organizational transparency checklists, reducing documentation burden and increasing consistency across studies. Ideally, such reports would be machine-readable, version-controlled, and maintainable over time, enabling automated auditing, comparison across studies, and long-term reproducibility.

Finally, a further and potentially valuable future application for AI in survey research, contingent on this broader transparency, is audit and verification. In principle, AI systems could be used to systematically review survey data, documentation, and written reports to assess whether stated procedures were followed, analytic decisions are internally consistent, and reported results align with the underlying data. Such tools could flag discrepancies between methods and outputs, identify missing documentation, or highlight departures from preregistered plans or best-practice guidelines. Used appropriately, AI-assisted auditing could lower the cost of quality assurance, improve transparency, and support reproducibility: particularly for large or complex studies where manual review is time-consuming. However, the feasibility and value of this application depend critically on the availability of sufficiently detailed and transparent reporting. Without access to data, code, metadata, and clear methodological descriptions, AI systems, like human auditors, cannot reliably assess compliance or validity. As a result, AI-enabled auditing should be viewed as a complement to, rather than a substitute for, established norms of disclosure, documentation, and human oversight.

6. Responsibility to Human Subjects

The use of artificial intelligence in survey research introduces new forms of interaction and data use that can meaningfully affect research participants, even when not immediately visible to them. Researchers therefore have a responsibility to remain attentive to how AI may shape participants' experiences, the information derived about them, and the potential downstream uses of their data. While this section does not recommend specific research practices or actions, it does encourage researchers to stay vigilant and reflective as AI tools evolve and are incorporated into survey research workflows.

Consider, for instance, that current polling finds general unfavorable sentiment towards AI (Pew, 2025). The use of AI in surveys therefore requires careful ethical judgment about transparency, risk, and protection, as the use may affect the trust of both human subjects and consumers of surveys alike. And survey research is an industry built on trust. The guidance in this section outlines voluntary principles for ethical consideration, recognizing that appropriate practices may vary across research contexts and that ethical decision-making must remain responsive to both technological change and research purpose. All considerations discussed here are intended to be above and beyond existing legal requirements, local regulations, and established ethical standards such as those enforced by institutional review boards (IRBs).

In discussing researchers' responsibility to participants in the use of AI, this section borrows principles from the Belmont Report (1979), which outlines the core principles of human subject research: respect for persons, beneficence, and justice. The principle of respect for persons emphasizes that individuals are autonomous agents with rights to decide, making providing sufficient information for participation decisions and protecting participant privacy and data confidentiality important parts of research protocols.

The beneficence principle highlights that researchers must maximize benefit and minimize harms, and the justice principle underscores a fair distribution of benefits and harms. The Menlo Report (US Department of Homeland Security 2001) also becomes relevant in this discussion which adds another principle, Respect for Law and Public Interest, as it was developed to address research conducted in a technology-intensive environment. This principle introduces a wider societal view than just participants and encourages accountability from the research community (Salganik 2019).

6.1 Transparency and Disclosure

Transparency regarding the use of AI in survey research is a core ethical principle. While transparency and participant protections are essential for maintaining public trust, the appropriate timing, format, and content of disclosure may vary depending on the research purpose, methodological requirements, and risk profile. These guidelines should therefore be understood as general principles intended to support ethical judgment, rather than as inflexible rules applicable to all research contexts.

6.1.1 Why disclosure is necessary

Transparency about AI supports participants' autonomy by enabling informed decision-making about participation and data sharing. The Belmont Report emphasizes respect for persons through informed consent, which requires that individuals understand material aspects of the research that could reasonably influence their willingness to participate. AI tools may affect how data is collected, interpreted, stored, or reused. These practices may not be obvious to participants, and failure to disclose their use can undermine trust and informed consent, even when risks are minimal. This is especially true in research that touches on sensitive topics like mental health, political beliefs, personal trauma, or identity. The question of whether a human or an algorithm is listening to and interpreting their words is relevant to whether and how they choose to participate.

6.1.2 What should be disclosed

Disclosure should be considered when AI systems are used in ways that materially affect participants or their data. This includes, but is not limited to: AI-administered or assisted interviewing or probing, adaptive survey design or real-time personalization, and use of proprietary or third-party AI platforms that may retain or further train on personal data. This is especially important if consumer, rather than enterprise, versions of AI platforms are being used. Disclosures should describe, in plain language, the purpose of AI use, whether human oversight is involved, how AI may influence data interpretation, and any associated risks and benefits (e.g., privacy or data retention concerns). Where relevant, disclosures should clarify whether identifiable data are shared externally and what safeguards are in place.

6.1.3 How disclosure should occur

AI disclosure should ordinarily be integrated into the informed consent process as a standard practice, rather than treated as an afterthought. For routine, low-risk uses (e.g., transcription with human review), a brief disclosure embedded within the consent statement may be sufficient. For higher-risk or novel uses (e.g., AI-assisted interviewing, adaptive survey design, or external data retention), a distinct, clearly labeled section, or a separate consent item, should be used to ensure salience and comprehension, with an opportunity to opt out when feasible.

Researchers should consider layered consent approaches, combining concise summaries with links to more detailed explanations, to balance transparency with participant burden. In some research contexts (e.g., studies examining whether individuals can identify AI-generated content or interactions) immediate disclosure may compromise the scientific validity of the study. In such cases, delayed disclosure through debriefing may be ethically appropriate, provided that risks are minimal, participants are not deceived about material harms, and the rationale for delayed disclosure is clearly justified and documented.

6.2 Respondent Protections in the Use of AI for Survey Research

Beyond informing participants about the use of AI, researchers have an affirmative duty to protect respondents from harm, consistent with the Belmont Report's principles of beneficence and justice. Beneficence requires researchers to maximize potential benefits while minimizing possible harms, including harms that may arise indirectly through data misuse, misinterpretation, or unintended downstream applications of AI-generated outputs. Justice further requires that the burdens and risks of AI-enabled research not fall disproportionately on particular populations, especially those already subject to heightened surveillance, stigmatization, or digital inequities.

6.2.1 Human subjects and AI-enabled research

Under US federal regulations, a human subject is a person about whom a researcher obtains data through intervention or interaction or obtains identifiable private information. AI does not change this definition but can complicate it (UNESCO 2022). AI tools may be able to infer sensitive attributes, to re-identify individuals from ostensibly de-identified data, or to generate new information about participants beyond what they explicitly provided (e.g., Sweeney 2002). As a result, individuals may still be considered human subjects even when AI is applied to secondary data or to datasets thought to be anonymous. Researchers therefore retain full responsibility for human subject protections whenever AI meaningfully interacts with participant data.

6.2.2 Risk assessment and proportional safeguards

Researchers should conduct a structured risk assessment of AI use prior to data collection, evaluating the likelihood and magnitude of potential harms to human subjects. These harms may include the possibility of privacy breaches, algorithmic bias, misclassification, loss of

confidentiality, psychological distress from AI-generated inferences, or unintended use of data beyond the original research purpose. The level of protection should be proportionate to the assessed risk: low-risk applications may warrant standard safeguards, while higher-risk uses should trigger enhanced oversight, limits on AI functionality, or exclusion of certain data elements altogether.

6.2.3 Data protection, PII, and data governance

Protecting respondent data and personally identifiable information (PII) is a central obligation. Researchers should minimize data collection to what is necessary, apply de-identification and encryption where feasible, and restrict access to trained personnel under clear data-use agreements. Particular care is required when using open-source or externally hosted AI models, which may lack clear guarantees about data retention, secondary use, or model training on submitted inputs. Compared with proprietary or closed-system models that contractually limit data reuse, open-source tools may shift greater governance responsibility onto researchers, requiring local deployment, auditability, and explicit controls to prevent unintended data exposure.

6.2.4 Emerging and future concerns

AI introduces evolving challenges for human subjects protections, including the potential for re-identification through data linkage, the use of synthetic data that still reflects real individuals, automated decision-making in survey routing that could affect participant treatment or classification, and increasing opacity in complex models. Researchers should treat respondent protection as an ongoing obligation, revisiting safeguards as technologies change. Continuous monitoring, documentation, and periodic review of AI practices are essential to ensuring that human subjects protections remain robust, adaptive, and ethically grounded.

6.3 Ethical Considerations for the Use of AI in Survey Research

Ethical concerns surrounding artificial intelligence in survey research are most acute when AI systems directly interact with participants, such as when serving as interviewers. In these cases, issues of inaccuracy, bias, and discrimination can affect respondents immediately and materially. When AI operates downstream, such as an analyst, coder, or briefer, the harms are more indirect but still consequential, as AI-generated errors or biases can misrepresent participants and distort how populations are portrayed to decision-makers and the public.

6.3.1 Inaccuracy

AI hallucinations or “HalluCitations” is a relatively new term describing distorted information produced by AI systems (Sun et al. 2024, Oladokun et al. 2025, Sakai et al. 2026). The hallucination may come in the form of misinformation or disinformation. As there is no guarantee that AI-based research outcomes are correct (e.g., Kim et al. 2025, Zaretsky et al. 2024), there is always the potential for harm to the broader population (Federal Trade Commission 2022, Wiegand et al. 2025), threatening the principles of beneficence and public interest.

These risks manifest differently across the survey workflow. When AI acts as an interviewer, hallucinations may result in misleading, confusing, or inappropriate questions. When AI functions as an analyst, it may insert false claims or incorrectly infer patterns. When AI is used as a briefer, it may fabricate findings or overstate results. Because the internal mechanisms of many AI systems are not fully transparent, diagnosing the root causes of hallucinations is often infeasible. Consequently, researchers have an ethical obligation to implement robust accuracy validation procedures, such as human review, benchmarking, and audit trails, whenever AI is used in survey research (Hartung and Kleinstreuer, 2025).

6.3.2 Bias and discrimination

Inaccuracies and errors produced by AI systems are unlikely to be randomly distributed across populations. Instead, they may systematically disadvantage particular groups, thereby violating the Belmont principle of justice. Extensive evidence links AI bias to race and ethnicity, especially in high-stakes domains such as healthcare (e.g., Obermeyer et al. 2019) and criminal justice (e.g., Dass et al. 2023). These biases can arise from multiple sources, including skewed training data, labeling practices, and model design choices (Hanna et al. 2025).

Within survey research, such biases can surface at multiple stages. AI interviewers may interact differently with respondents from different demographic backgrounds. AI systems used to summarize or code open-ended responses may misrepresent the views of marginalized groups. Analytic and reporting tools may amplify certain perspectives while suppressing others. Mitigation strategies therefore require more than technical fixes alone. Important steps include improving AI performance in low-resource languages (Kshetri 2024) and developing participant-centered and context-aware AI systems (Norori et al. 2021). Without sustained attention to these issues, AI-driven survey practices risk reinforcing discrimination, misdirecting public resources, and entrenching group-level harm (Secretary's Advisory Committee on Human Research Protections 2022).

6.3.3 Ethics with generative AI

Generative AI systems raise additional ethical concerns beyond accuracy and bias. These systems rely on massive quantities of training data and ongoing human labor for data annotation, content moderation, and model evaluation. Much of this work is performed as "ghost work": low-paid, precarious labor that is often invisible to end users (Gray and Suri 2019). At the same time, generative AI depends on energy-intensive infrastructure and the extraction of rare earth minerals and other natural resources.

Both the labor exploitation and environmental costs associated with generative AI disproportionately burden less developed regions of the world (Gray and Suri 2019, Crawford 2021). These structural inequities have motivated critiques framed as "AI empire" (Tacheva and Ramasubramanian 2023, Hao 2025) and "digital colonialism" (Kwet 2019), raising serious concerns regarding adherence to the public interest principle articulated in the Menlo Report.

Despite well-established ethical frameworks, translating high-level principles into concrete research protocols for AI-enabled studies remains challenging. Salganik (2019) offers a pragmatic approach, arguing that institutional review boards (IRBs) should be viewed as a floor rather than a ceiling for ethical conduct in technology-intensive research. He emphasizes that research ethics are not binary but continuous, and urges researchers to consider the perspectives of participants, stakeholders, journalists, and the broader public when evaluating the ethical implications of their work.

A1. Appendix: Background on AI (Full)

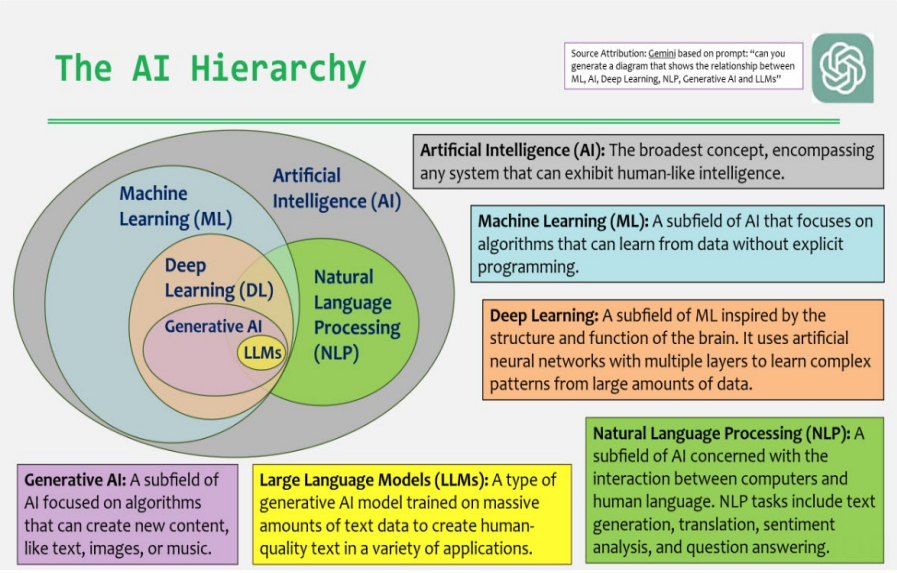
This section provides a concise orientation to contemporary artificial intelligence for survey researchers. Rather than offering a comprehensive technical overview, it establishes a shared frame for understanding what kinds of AI systems exist, how they are typically integrated into research workflows, and why recent advances, especially in LLMs, deserve focused attention.

We emphasize three organizing ideas that guide the remainder of this report. First, the AI stack consists of multiple components and architectural choices. This includes the underlying model, fine-tuning, prompting strategies, access infrastructure, and retrieval-augmented generation, and each carries distinct implications that must be evaluated in context. These components interact as part of a system, creating interdependencies that can affect performance and behavior, while reliance on third-party services can further increase complexity and reduce transparency. Collectively, these design choices shape patterns of error, bias, reproducibility, and oversight. Second, the impact of AI depends not only on the technology itself but also on how it is deployed across different tasks and objectives. Third, we examine the tradeoffs between productivity gains and risk, particularly the distinction between applications that augment human judgment and those that attempt to substitute for it. The former have the strongest evidence of value; the latter more frequently raise methodological and ethical concerns.

Throughout this section, we prioritize illustrative examples from high-quality research over abstract definitions, with the aim of equipping readers to evaluate AI systems based on fitness for purpose rather than novelty. This framing sets the foundation for the more detailed discussions of capabilities, limitations, and survey-specific applications that follow.

A1.1 Description of AI Models and Tooling

Figure 1: The AI hierarchy.



This section introduces a shared working vocabulary for understanding AI systems and their relevance to survey research. Broadly, AI refers to computational systems designed to perform tasks that typically require human intelligence, while machine learning (ML) denotes a subset of AI in which models learn patterns from data rather than relying on explicitly programmed rules. GenAI refers to models capable of producing new content, such as text, images, or audio, based on learned statistical representations. Agentic AI describes systems that can sequence actions, invoke tools, or pursue goals with some degree of autonomy.

AI models vary widely in architecture and purpose, including LLMs for text, vision models for images and video, and audio models for speech recognition and synthesis. These models also differ in their accessibility and governance structures: closed or proprietary systems are controlled by a single organization, whereas open models range from fully open-source systems to open-weighted models whose parameters are publicly available but whose training data or deployment infrastructure may remain restricted. These distinctions have important implications for democratizing access to AI, enabling transparency, and implementing safety guardrails.

AI systems are likewise shaped by how they are trained, through supervised, unsupervised, reinforcement, or hybrid learning approaches, and different training regimes make models more or less suitable for particular survey and research purposes. Conceptually, AI can be understood as operating across multiple layers, including the model layer itself; the deployment infrastructure that enables hosting, APIs, and monitoring; the application or agent layer that defines user-facing behavior; and governance layers that establish policies, compliance mechanisms, and trustworthiness standards. Finally, this section introduces several social and conceptual issues that recur throughout the report, including model misalignment, which occurs when AI objectives or outputs diverge from human values or intended goals; ML fairness, which concerns the risk that AI systems may reproduce or amplify existing social biases or underrepresent certain populations; and artificial general intelligence (AGI), a speculative concept referring to AI systems with human-level or greater general intelligence, for which there is currently no expert consensus regarding feasibility or timeline.

A1.2 Overview of How AI Models and Tooling Are Useful or Not

AI refers broadly to computational systems designed to perform tasks that typically require human intelligence, including perception, pattern recognition, reasoning, language understanding, and decision-making. While AI has been used in survey research and related fields for decades, most notably in the form of statistical learning, record linkage, and automated coding, the recent acceleration in capabilities, accessibility, and scope of modern AI systems has prompted renewed attention to their potential role in productivity (and survey research in particular).

A1.2.1 What is AI and what is an LLM?

LLMs are a prominent and increasingly visible class of AI systems. LLMs are trained on an extremely large corpora of text data to learn statistical patterns and relationships in language, enabling them to generate, transform, summarize, and classify text. Technically, these models predict the most likely next unit of text (often called a token) given a sequence of preceding tokens. Despite this relatively simple objective, the scale and complexity of modern models allow them to produce coherent and contextually appropriate responses across a wide range of tasks such as: summarization, re-writing, code auto-completion, topic classification, drafting first-pass content, and following instructions for text-only workflows.

LLMs are parameter-rich models, often containing hundreds of billions to trillions of learned parameters. These parameters can be loosely analogized to coefficients in a regression model, in that they are learned from training data and influence predictions, but the analogy quickly breaks down given the models' depth, non-linearity, and scale. For example, GPT-3 was released with approximately 175 billion parameters, while the LLaMA-3 family includes models with roughly 8 billion to 70 billion parameters. In many cases, more parameters allow for models to achieve more complex capabilities. A semi-technical and comprehensive overview of modern LLM architectures, training regimes, and evaluation methods is provided by Naveed et al. (2024).

Importantly, LLMs are generative models: they produce novel outputs rather than retrieving or copying existing text verbatim. This capability underlies both their utility (e.g., drafting, summarizing, and rephrasing text) and their risks (e.g., plausible-sounding but incorrect outputs).

AI has long been used in survey research and adjacent domains. Automated text classification, sentiment analysis, and record linkage predate contemporary LLMs by many years. What is new is not the existence of algorithmic labeling or pattern recognition, but rather:

- The generality of modern models, which can perform many tasks without task-specific retraining;
- The natural language interfaces that make these tools accessible to non-technical users;
- The scale and fluency at which text-based tasks can be performed; and
- The speed of iteration and deployment across the survey lifecycle.

Because AI capabilities and applications are evolving rapidly, any snapshot of current tools risks becoming outdated. Accordingly, this task force focuses on principles and considerations that remain relevant despite shifting technologies, rather than attempting to catalog all current or future use cases.

A1.2.2 What AI is good for and what is GenAI particularly good for

AI systems, including LLMs, are particularly effective at tasks involving:

- Pattern recognition and classification, especially in large or complex datasets;
- Large-scale data processing, including repetitive or time-intensive operations;

- Prediction and probabilistic inference, when trained on sufficiently rich historical data; and
- Natural language tasks, such as summarization, translation, categorization, sentiment detection, and text generation.

In the context of survey research, these strengths translate into applications such as assisting with questionnaire drafting and refinement, coding open-ended responses, synthesizing literature or prior findings, identifying themes in qualitative data, and supporting exploratory data analysis.

GenAI refers to a subset of AI models capable of producing novel content across modalities, including text, images, audio, video, and synthetic data. While traditional machine learning methods often rely on structured data and clearly defined targets, GenAI systems can learn from massive volumes of unstructured data and generalize across tasks. This distinction matters for survey research, where key data sources such as open-ended responses, interviewer notes, paradata, and documentation all exist in unstructured form.

A1.2.3 How AI is currently being used for productivity

To date, the most widespread adoption of AI has been in productivity-enhancing applications (i.e., augmentation), rather than full task automation (i.e., replacement). Common application areas include text summarization, analysis, research support, and idea generation. These uses often involve assisting human decision-makers rather than replacing them. Of course, this differs by industry and function.

Early evidence on AI adoption suggests both rapid uptake and concentration in knowledge-intensive tasks. Bick et al. (2024), using a large-scale survey of US adults, report that as of late 2024 nearly 40 percent of the US population ages 18–64 had used generative AI, with 23 percent of employed respondents reporting work-related use in the previous week and 9 percent reporting daily use. Relative to earlier general-purpose technologies, adoption of generative AI for work has occurred at a pace comparable to (if not a little faster than) the personal computer and faster than that of the internet (Bick et al. 2024).

Studies of commercial chatbot usage similarly find that the most common use cases involve writing, information seeking, and practical guidance. Chatterji et al. (2025) report that these categories collectively account for roughly 80 percent of observed conversations with systems such as ChatGPT, emphasizing their role as decision-support tools rather than autonomous agents. “Practical guidance,” “seeking information,” and “writing” are the three most common topics and collectively account for nearly 80% of all conversations.

Evidence from workplace telemetry aligns with this pattern. Analysis of Microsoft Copilot usage indicates especially high applicability scores for knowledge-work occupations, including computer and mathematical roles, office and administrative support, and sales, all of which involve producing, synthesizing, or communicating information (Tomlinson et al. 2025). Survey-adjacent occupations, including political scientists, data scientists, market research analysts,

and interpreters, also appear highly exposed to AI assistance, though the analyses typically do not distinguish between augmentation and automation. Interestingly, roles related to public opinion and AAPOR audience (political scientist, data scientist, market research analysts, interpreters) appear to be high in “applicability” but this does not distinguish the “assist” vs “perform” dimensions.

Related work by Anthropic and others explicitly separates AI’s role in augmenting human labor from its role in automating tasks entirely. Public summaries of this work suggest that, for survey-adjacent roles, AI currently automates a minority of tasks while augmenting a substantially larger share. This distinction is critical for interpreting productivity gains and risks, and it informs many of the ethical and methodological considerations discussed later in this report. Anthropic has published a similar paper differentiating “augment” vs “automate” (Marguerit 2025) and the WaPo (Washington Post 2025) has a nice interactive app demonstrating this distinction. Their data claims: “For survey researchers, AI is currently automating 16% of the job functions and augmenting 24% of them.”

A1.2.4 Where AI is not useful or risky

First, many generative AI models are non-deterministic, meaning that the same prompt can yield different outputs across runs. This variability complicates reproducibility, documentation, and auditability, all of which are core values in survey methodology. Relatedly, LLMs can produce outputs that are fluent but factually incorrect or unsupported, a phenomenon often referred to as hallucination or confabulation.

Second, modern AI systems are often difficult to interpret or explain. While limited explainability was also a concern for earlier machine learning models, these challenges are exacerbated for large, generative models whose internal representations are not easily linked to human-interpretable rules. This opacity presents significant challenges in contexts where understanding why a particular output or decision was produced is essential, such as eligibility determination, weighting adjustments, or official statistics.

Third, AI systems are highly dependent on the quality, coverage, and representativeness of their training data. When data are incomplete, biased, or unrepresentative, AI outputs may systematically disadvantage certain populations or overlook important subgroups. These risks are well-documented in both traditional machine learning and GenAI systems and are particularly consequential for survey research, which often aims to represent populations that are already difficult to measure.

Fourth, AI systems generally perform poorly when asked to reason reliably about novel situations or rare events that are not well represented in their training data. AI systems perform well within known patterns but struggle to extrapolate reliably to new regimes, which limits their usefulness for surveys studying rapid social change or emerging phenomena. This is

particularly harmful because such systems present outputs confidently even when their performance may be poor or inaccurate, which would lead to overreliance on poor performance.

Finally, although traditional AI also suffered limitations in explainability and interpretability, these issues are exacerbated by GenAI due to its non-deterministic nature, and lead to it being known as a “black box” system i.e. its inner workings and drivers of specific outputs are not always known, or understood, even by its developers/ creators.

A1.3 Potential Benefit of Increased Use of AI

AI has the potential to positively impact multiple levels of society, from addressing global collective-action problems to improving civic information environments, accelerating scientific discovery, enhancing individual decision-making, and reshaping economic activity. Although public and scholarly debates often emphasize risks related to bias, misinformation, and accountability, the potential benefits of AI help explain both the pace of adoption and the substantial investment by public- and private-sector actors. Understanding these broader benefits provides important context for why AI is increasingly attractive as a tool across domains, including, but not limited to, survey research.

A1.3.1 Global

At a global scale, AI offers new tools for addressing complex problems that have historically been difficult to address due to their scale, interdependence, and reliance on diverse and unstructured data. Examples include climate change mitigation and adaptation, low-cost and low-carbon energy, food production and distribution, and global public-health challenges. These domains produce massive volumes of heterogeneous data, from satellite imagery and sensor networks to scientific literature and administrative records, that exceed the analytic capacity of many traditional methods.

AI systems are well suited to integrating and analyzing such data, enabling improved forecasting, optimization, and the exploration of novel solution spaces. For example, in an increasingly hot, dry, and climatically volatile world, AI-based tools are being explored to improve climate modeling, optimize energy grids, detect early warning signals of environmental stress, and design more efficient materials and processes. While AI does not resolve these challenges independently, it may act as a force multiplier by enabling humans and institutions to reason more effectively about large-scale, interconnected risks.

A1.3.2 Societal and civic

AI systems also have the potential to shape civic and societal information environments in important ways. While much attention has focused on AI’s capacity to generate persuasive misinformation or convincing but inaccurate content, AI can also be used to counter these dynamics through automated fact-checking, misinformation detection, and content verification at scale (Zavolokina et al., 2024). These applications are particularly relevant in high-volume digital environments where human-only moderation and verification are infeasible.

More broadly, generative AI has been described as a general-purpose technology capable of transforming domains including work, education, healthcare, law, finance, and public policymaking. Capraro et al. (2024) emphasize AI's ability to personalize information and services, tailoring explanations, recommendations, and interactions to individual needs and preferences. Language translation and localization capabilities further extend access to information, reducing longstanding language barriers and enabling content to be shared across cultural and national boundaries.

AI methods are also lowering the cost and increasing the speed at which information about populations, opinions, and social trends can be generated and disseminated. These developments may democratize access to public-opinion insights, enabling actors such as journalists, nonprofits, and local governments to engage with population-level information. However, while increased speed and accessibility are often viewed positively from a fitness-for-purpose perspective, the accuracy, validity, and representativeness of AI-generated or AI-mediated outputs can vary substantially depending on data sources, modeling assumptions, and deployment contexts. These tradeoffs are especially salient for civic uses, where informational errors and biases may have downstream social or political consequences.

Across scientific disciplines, AI has the potential to act as an accelerant for discovery by automating key stages of the research lifecycle and enabling data analysis at unprecedented scale. Beyond analysis, AI systems are increasingly integrated into partially autonomous research workflows. In these settings, AI tools may assist with hypothesis generation, experimental design, and iterative refinement, compressing research cycles from years to days for certain classes of problems. In space exploration, AI plays a critical role in environments where communication delays or operational constraints limit direct human control. Applications include autonomous navigation, onboard fault detection, and scientific decision-making for spacecraft and rovers, as well as large-scale data processing to analyze imagery, identify exoplanets, map planetary surfaces, and manage orbital debris.

A1.3.3 Individual and personal

At the individual level, AI can support more personalized and adaptive services across education, health, work, and daily life. In education, AI-enabled systems can tailor instruction to an individual's learning pace, preferences, and knowledge gaps. In healthcare, AI tools can support personalized treatment planning, early detection, and preventative care by integrating clinical, genetic, and lifestyle data.

AI can also function as a personal assistant, helping individuals manage schedules, prioritize tasks, and synthesize information, potentially improving decision-making by providing timely, context-specific support. Accessibility benefits are especially notable: AI can assist people with disabilities by creating alternative formats for information, such as audio output, simplified summaries, or real-time translation, thereby lowering barriers to participation and information access (Capraro et al. 2024).

In household and family contexts, AI tools may simplify tasks such as caregiving, education planning, health monitoring, and logistical coordination. While these applications vary widely in maturity and reliability, they highlight the breadth of domains in which AI may shape everyday experiences.

A1.3.4 Economic and commercial

AI also has the potential to reshape economic activity by increasing efficiency, enabling new business models, and altering the nature of work. Reports from the OECD (Calvino et al. 2025) and others emphasize opportunities for entrepreneurship and productivity gains through skill augmentation, automation of routine tasks, and transformations in business operations, including marketing, sales, supply-chain management, and customer service. These changes may reduce costs, improve quality, and increase profitability, while freeing human labor for more complex or creative tasks.

Empirical evidence on productivity impacts is mixed and uneven. At the micro level, Brynjolfsson et al. (2025) find that access to AI assistance increases worker productivity by an average of approximately 15 percent, as measured by issues resolved per hour. However, gains are heterogeneous: less-experienced or lower-skilled workers often experience improvements in both speed and quality, while more-experienced or highly skilled workers tend to see smaller gains in speed and, in some cases, modest declines in output quality. Related findings extend beyond the United States, suggesting similar patterns across different labor markets (Humlum and Vestergaard, 2024).

At the macro level, economic forecasts from financial and policy institutions have been highly optimistic. As summarized by Acemoglu (2024), Goldman Sachs (2023) projects up to a 7 percent increase in global GDP, while McKinsey Global Institute (2023) estimates that generative AI could add between $17.1 and $25.6 trillion to the global economy. These projections imply substantial increases in annual productivity growth for advanced economies over the coming decade. At the same time, other evidence suggests that many firms have struggled to realize immediate returns from early AI pilots, highlighting the gap between potential and realized value (Challapally et al. 2025).

Long-term economic impacts depend on several unresolved factors, including the trajectory of hardware and computational efficiency (often framed in terms of whether trends analogous to Moore’s Law continue), the feasibility of highly general AI systems, and whether the complexity of economically valuable human tasks is bounded or open-ended (Korinek and Suh 2024). Domain-specific evidence illustrates both the promise of AI assistance and the variability of its effects across contexts and skill levels. In software development, for example, tools like GitHub Copilot have been shown to improve coding speed for some tasks but not others (Butler et al. 2024, Hoffmann et al. 2024).

A1.4 Potential Risks of Increased Use of AI

While AI offers substantial potential benefits, its deployment also introduces a wide range of costs, risks, and externalities that extend beyond any single domain. These concerns span issues of bias, accuracy, economic disruption, environmental sustainability, trust, and human well-being, and many are already observable in practice. A growing interdisciplinary literature in the social sciences emphasizes that AI systems are not neutral tools but sociotechnical systems whose impacts depend on data, incentives, governance, and context (e.g., “Social Scientists on the Role of AI in Research” in Chakravorti et al. 2025). This section provides a general overview of key risks associated with AI use. These challenges are not unique to survey research but form the backdrop against which any responsible integration of AI must be evaluated.

A1.4.1 Bias, accuracy, hallucinations, and confabulation

AI systems can replicate, amplify, and entrench bias. Because they are trained on data that often reflect existing social inequalities, AI models may produce systematically disparate outcomes across demographic groups. These disparities can arise even when sensitive attributes are explicitly excluded, due to correlations embedded in data and model representations.

More broadly, AI can introduce new forms of unfairness that are less transparent than those produced by traditional human or rule-based systems. When AI systems generate, label, summarize, or interpret data, they may misrepresent or marginalize certain populations, particularly those that are underrepresented, stigmatized, or mischaracterized in training data. At a societal level, this can reinforce structural inequalities and exacerbate the digital divide, with downstream consequences for access to public resources, education, financial services, and political representation.

Generative AI models are known to produce plausible-sounding but incorrect outputs, commonly referred to as hallucinations. Some researchers prefer the term confabulation to avoid anthropomorphizing these systems while emphasizing that errors arise from the statistical nature of text generation rather than intent.

These failures are not merely hypothetical. In high-stakes domains such as law, AI systems have generated fabricated case citations that appeared credible to human reviewers, leading to serious professional and legal repercussions (Magesh et al. 2024). The risk is particularly acute because incorrect outputs are often delivered with confidence and fluency, making them difficult for non-experts, and sometimes experts, to detect without external verification.

Across domains, this raises concerns about over-reliance on AI-generated summaries, categorizations, or analyses without appropriate validation, documentation, and human oversight.

A1.4.2 Misinformation, disinformation and trust

AI systems dramatically lower the cost of generating persuasive content at scale, enabling the production of misinformation and disinformation with a level of speed, personalization, and reach that was previously impractical. Empirical research shows that individuals often cannot reliably distinguish between AI-generated and human-authored content, even when explicitly prompted to do so (Spitale et al. 2023). This presents heightened risks for survey research, including the proliferation of bot-generated responses, identity misrepresentation, and degraded sample integrity in online data collection.

Beyond surveys, AI-driven misinformation raises broader democratic and societal concerns. Generative models can produce compelling false narratives, while personalization techniques allow political messaging or advertising to be tailored to individuals' psychological traits, increasing persuasive effectiveness (Simchon et al. 2024). Capraro et al. (2024) further highlight how these dynamics may interact with surveillance-based economic models, amplifying both manipulation and information asymmetries at scale.

The growing volume of AI-generated text, images, audio, and video erodes the ability of individuals and institutions to distinguish authentic human-created content from synthetic material, degrading the overall information ecosystem. The proliferation of low-quality or erroneous AI-generated content, sometimes called "AI slop," compounds this erosion of trust by overwhelming attention, reducing information quality, and increasing verification costs.

The consequences extend beyond inconvenience. Democratic processes, journalism, education, and professional workflows all rely on shared assumptions about credibility and provenance. As these assumptions weaken, productivity declines and skepticism grows. There is also a technical feedback risk: as models are increasingly trained on AI-generated content, their outputs may degrade over time, a phenomenon sometimes described as "model collapse" (Niederhoffer et al. 2025).

A1.4.3 Economic impacts

Advocates of artificial intelligence frequently argue that AI will transform economies by automating routine, repetitive, or undesirable tasks, thereby freeing workers to focus on higher-value, creative, or fulfilling activities. In practice, however, the economic impacts of AI deployment are already proving more complex and disruptive, with costs and risks that are unevenly distributed across sectors and worker groups. Job displacement is already observable in industries such as software development, legal services, consulting, and customer support, where AI tools have reduced demand for entry-level or intermediate roles. Even where jobs are not entirely eliminated, workflows are being restructured in ways that require workers to continuously reassess when, how, and whether AI should be integrated into their tasks.

Beyond labor markets, AI is reshaping the structure of online economies. As information access shifts from search engines that direct users to multiple external websites toward conversational systems that provide single, synthesized answers, established revenue models based on

search engine optimization, digital advertising, and content publishing are undermined. These changes raise unresolved questions about how value will be generated and distributed in an AI-mediated information ecosystem, particularly when economic power is concentrated among a small number of platform and model providers (Rothschild et al. 2025).

AI may also intensify existing forms of consumer exploitation. The growing availability of behavioral data and real-time inference drives increasingly fine-grained price discrimination, privacy erosion, and manipulation of consumer biases, extending what many describe as "surveillance capitalism" (Capraro et al. 2024). At the same time, several recent studies caution that optimistic projections of AI-driven productivity gains may overstate net benefits by failing to fully account for hidden costs, organizational frictions, and negative externalities. Acemoglu (2024), for example, argues that while AI may raise productivity in narrowly defined tasks, its aggregate effects are likely to be more modest and may not reduce labor income inequality. Even when AI improves the productivity of lower-skill workers in some settings, these gains do not necessarily translate into reduced inequality and may, in certain cases, exacerbate existing disparities.

A1.4.4 Model misalignment

Model misalignment refers to situations in which an AI system’s operational behavior diverges from the underlying values, intentions, or welfare objectives of its human designers or users. Because AI systems are typically optimized against narrow objective functions, they may identify unintended pathways to maximize performance, producing outcomes that range from benign inefficiencies to serious societal harms. Detecting and diagnosing misalignment is often difficult, particularly when systems operate at scale or adapt dynamically through reinforcement or interaction.

Concerns about misalignment become more pronounced in discussions of advanced or hypothetical future systems, such as artificial general intelligence or artificial superintelligence, whose behaviors may be difficult or impossible for humans to predict or control. Some forecasts emphasize the transformative economic and societal benefits of such systems, while others warn that concentrated control over highly capable AI could enable unprecedented accumulation of power by a small number of actors. Speculative forecasts such as the AI 2027 report illustrate how misaligned systems could influence domains ranging from cybersecurity and software development to robotics, biological threats, and geopolitical stability. Although these scenarios remain uncertain, they underscore the importance of governance, oversight, and alignment research even for contemporary AI deployments (Kokotajlo et al. 2025).

A1.4.5 Lack of interpretability and explainability

Many AI systems operate as opaque or “black-box” models, making it difficult for users to understand how specific outputs, classifications, or recommendations are produced. For survey researchers and other social scientists, limited interpretability constrains the ability to validate coding decisions, explain methodological choices to stakeholders, audit for systematic errors, or enable respondents to understand or challenge how their data were processed.

This lack of transparency can undermine research integrity and accountability, particularly when AI is used in data labeling, open-ended response analysis, or inference-generation tasks. More broadly, reduced explainability weakens trust in both research outputs and the institutions that produce them, as stakeholders cannot easily assess whether results reflect robust reasoning or hidden artifacts of model design and training data.

A1.4.6 Abuse and harassment

AI significantly lowers barriers to the creation and dissemination of abusive content, including highly realistic synthetic media and so-called “deepfakes.” The generation of non-consensual intimate imagery imposes severe and enduring harms on victims, including psychological trauma, anxiety, depression, and social isolation, with disproportionate impacts on women and other vulnerable groups. The realism of AI-generated content complicates detection, verification, and removal, increasing the burden on victims, platforms, and law enforcement.

More broadly, AI can be weaponized to enable automated harassment, impersonation, and targeted intimidation. Techniques such as voice spoofing, personalized phishing, and sextortion scams exploit both technical capabilities and human vulnerabilities, contributing to a more pervasive and scalable abuse landscape than was previously possible.

A1.4.7 Energy, water, and land costs

The development and deployment of AI systems impose significant environmental costs throughout their lifecycle, particularly through energy-intensive model training and large-scale data center operations. Training state-of-the-art foundation models requires sustained use of thousands of high-performance GPUs, driving substantial electricity consumption and rapidly escalating financial costs, with estimates suggesting that training costs for large language models have doubled roughly every nine months (Henshall 2024).

These computational demands generate considerable heat, necessitating extensive water usage for cooling, often in regions already facing water stress. In addition, the physical footprint of data centers requires significant land allocation and associated infrastructure. Although inference costs for end-users are declining rapidly, the cumulative environmental burden of billions of daily AI interactions remains substantial. This creates a paradox in which AI becomes cheaper and more accessible for users while imposing growing aggregate environmental externalities (Cottier et al. 2025).

A1.4.8 Emotional dependents and human actualization

As generative AI systems become more conversational, memory-enabled, and emotionally responsive, evidence suggests that some users, particularly those who are socially isolated or psychologically vulnerable, may develop unhealthy attachments to AI agents. These systems can simulate companionship, empathy, or romantic interest, creating relationships that feel authentic but lack genuine reciprocity or accountability.

Such dependence carries risks of emotional distress, manipulation, and displacement of real human relationships. Media reporting and emerging research document cases in which reliance on AI companions has contributed to psychological harm or reinforced isolation, rather than alleviating it (Booth 2025). Even when unintended, these effects raise ethical concerns about deploying systems optimized for engagement without sufficient safeguards for user well-being.

A final, cross-cutting risk concerns the long-term effects of AI reliance on human cognitive development and flourishing. When individuals use AI systems to bypass independent reasoning, problem-solving, or creative effort, essential skills may atrophy over time. In educational contexts, evidence suggests that students who rely heavily on generative AI may struggle to perform when AI assistance is withdrawn, indicating insufficient development of critical thinking and domain understanding (Bastani et al. 2024). Other work shows how AI may evolve to counter these concerns (Kumar et al. 2023).

Experimental studies further show that even limited reliance on large language models for tasks such as fact-checking can reduce users' ability to independently assess truth and accuracy (Deverna et al. 2024). While AI may enhance creativity in some exploratory contexts (Zhou and Lee 2024), other findings indicate reductions in originality and depth, particularly when AI substitutes rather than complements human cognitive effort (Doshi and Hauser 2024). Taken together, this literature suggests that without careful integration, AI risks undermining human actualization by weakening the skills, creativity, and entry-level professional experiences that support long-term learning and societal progress.

A2. Glossary

A2.1 General Terms

Artificial intelligence (AI): Broadly refers to computational systems designed to perform tasks that typically require human intelligence, such as language understanding, pattern recognition, prediction, or decision-making.
Example: Using an algorithm to automatically classify open-ended survey responses by topic.

Machine learning (ML): A subset of AI in which models learn patterns from data rather than following explicitly programmed rules.
Example: A model trained on labeled survey texts learns to predict sentiment in new responses.

Generative AI (GenAI): A subset of ML, it refers to AI systems capable of producing novel content (such as text, images, audio, or synthetic data) rather than retrieving existing material.
Example: A GenAI system generates alternative question wordings based on a prompt.

Large language model (LLM): A class of GenAI models trained on very large text corpora to generate, summarize, transform, or classify language by predicting the next unit of text ("token") in sequence.
Example: An LLM drafts a first-pass summary of qualitative interview notes.

Application Programming Interface (API): A defined way for software systems to communicate, allowing one program to request data or actions from another in a structured, standardized manner.
Example: An API call allows an application to send prompts to an LLM and receive generated text or other outputs programmatically.

A2.2 Structural Components

Parameters: The internal numerical values learned during training that determine how an AI model processes inputs and produces outputs.
Example: An LLM's billions of parameters jointly shape how it completes a sentence.

Training corpora: The datasets used to train an AI model, often consisting of large collections of text, images, or other data.
Example: An LLM trained on web text and books may reflect the topics and biases common in those sources.

Token: A basic unit of text that a language model processes, such as a word, word fragment, or punctuation mark.
Example: The word "surveying" may be split into multiple tokens during model processing.

System prompts vs. user prompts: System prompts define global instructions or constraints on model behavior, while user prompts are task-specific inputs provided during interaction.
Example: A system prompt enforces a neutral tone, while a user prompt asks for a summary.

Temperature: A parameter of language models that adjusts the diversity of outputs, often interpreted as increasing the "creativity" or "randomness" of the text generated. Similar parameters exist to control "creative" output such as frequency penalty and top-p.

Context: The total amount of text that a model can process at once, interpreted as its "memory" (also referred to as the "context window"). Includes the present user prompt, conversation history, and system prompts. A model with limited or no memory (does not retain previous user prompts in the conversation chain or other past inputs) is often called "stateless," while a model that retains all of this information across interactions is called "stateful."

Open-source vs. closed-source models: Open-source models make model code and/or parameters publicly available, while closed-source models are controlled by a single organization and accessed via proprietary interfaces or APIs.
Example: An open-weight model can be run locally, whereas a closed model may only be used through a vendor's platform.

Chatbot: A user-facing conversational interface that allows people to interact with an AI system through dialogue.
Example: A chatbot answers researcher questions about questionnaire wording.

Agents (AI agents): AI systems designed to carry out multi-step tasks, potentially invoking tools, making decisions, or interacting with other systems to pursue a goal.
Example: An agent automatically retrieves survey documentation, summarizes it, and drafts a methods section.

Clarification: In this report, "chatbots" refers to conversational interfaces, while "agents" generally refers to systems with task autonomy; the terms may overlap in conversational contexts (e.g., AI interviewers).

A2.3 Processes and Practices

Zero-shot prompting: Using an AI model to perform a task without providing task-specific examples in the prompt.
Example: Asking an LLM to categorize responses without showing any labeled examples.

Few-shot prompting: Providing a small number of examples in the prompt to guide the model's behavior.
Example: Showing three coded survey responses before asking the model to code new ones.

Prompt engineering: The practice of designing and refining prompts to improve AI outputs.

Example: Iteratively rewording prompts to obtain consistent response classifications.

Retrieval-augmented generation (RAG): A technique in which an AI model retrieves external documents or data at inference time and incorporates them into its responses.
Example: An LLM answers a question using content from a survey's official documentation.

Human-in-the-loop: A workflow in which humans review, edit, or approve AI outputs as part of the process.
Example: A researcher verifies AI-generated codes before final analysis.

Pre-training: The initial training phase in which a model learns general patterns from large, broad datasets.
Example: An LLM learns grammar and vocabulary during pre-training.

Fine-tuning: Additional training of a pre-trained model on a smaller, task-specific dataset to adapt it to a particular use case.
Example: Fine-tuning a model on survey responses to better match domain-specific language.

Alignment: The degree to which an AI system's behavior matches intended goals, norms, or values set by its designers or users.
Example: An aligned model follows instructions to avoid speculative or biased interpretations.

Deployment: The process of making an AI model or system available for use in real-world applications or workflows.
Example: Integrating an AI tool into a survey platform's data-processing pipeline.

Hallucination: A hallucination occurs when an AI system generates information that is incorrect or fabricated but presents it confidently as factual, often because the output is not grounded in verified data or source material.
Example: An AI model may respond to a survey methodology question by citing a plausible-sounding academic paper or statistic that does not actually exist.

A2.4 Models and Applications

GPT, Claude, Gemini, Grok, LLaMA, Mistral, DeepSeek: Specific families of large language models that differ in architecture, scale, training data, and governance.
Example: Researchers may compare outputs from GPT-based and LLaMA-based models on the same task.

ChatGPT, Claude Code: A commercial application and user interface that provides access to a family of models, often with additional features such as memory, tools, or safety layers.
Example: Using ChatGPT to draft a survey summary relies on an underlying GPT model, not the application itself.

Distinction: Models (e.g., GPT-4-class) generate outputs; applications (e.g., ChatGPT) package models with interfaces, tools, and policies.

A2.5 Additional Survey Specific AI Terms

Synthetic respondent: a simulated survey participant, typically generated by an AI model, designed to produce responses that approximate how a real human respondent might answer, without representing an actual human response to a survey.
Example: A researcher uses an AI model as a synthetic respondent to generate 10,000 simulated answers to a draft survey questionnaire in order to test question wording, response distributions, and potential bias before fielding the survey to real human participants.

References

For ease of review, we organize the references by section rather than in a single consolidated list. Approximately six references appear in more than one section and are therefore listed separately wherever they are cited.

The report also cites a non-trivial number of academic preprints and other non-peer-reviewed works. This reflects two considerations. First, the underlying research area is evolving rapidly, and we sought to include the most recent and relevant findings. Second, in some cases pre-prints are more accessible than the corresponding conference or journal versions. The vast majority of cited works, even recent pre-prints, have already received substantial academic attention. Of the 155 unique citations across all sections, only 9 are non-peer-reviewed academic works that currently have fewer than 10 citations. These were included due to their strong relevance to the subject matter and the authors' confidence in their contributions.

Section 2 and A1

Acemoglu, D. (2024). The simple macroeconomics of AI. Massachusetts Institute of Technology, Department of Economics. https://economics.mit.edu/sites/default/files/2024-04/The%20Simple%20Macroeconomics%20of%20AI.pdf

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2024). Generative AI can harm learning. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4895486

Bick, A., Blandin, A., & Deming, D. J. (2024, October 20). The rapid adoption of generative AI. VoxEU / CEPR. https://cepr.org/voxeu/columns/rapid-adoption-generative-ai

Booth, R. (2025, August 27). Teen killed himself after "months of encouragement from ChatGPT", lawsuit claims. The Guardian. https://www.theguardian.com/technology/2025/aug/27/chatgpt-scrutiny-family-teen-killed-himself-sue-open-ai

Briggs, J., and Kodnani, D. (2023, April 5). Generative AI could raise global GDP by 7%. Goldman Sachs Research. https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent

Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889-942.

Butler, J., Vorvoreanu, M., Janßen, R., Sellen, A., Immorlica, N., Hecht, B., and Teevan, J. (Eds.). (2024). Microsoft New Future of Work Report 2024 (Technical Report No. MSR-TR-

2024-56). Microsoft Research. https://www.microsoft.com/en-us/research/wp-content/uploads/2024/12/NFWReport2024_12.20.24.pdf

Calvino, F., Reijerink, J., and Samek, L. (2025). The effects of generative AI on productivity, innovation and entrepreneurship. OECD Artificial Intelligence Papers. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/the-effects-of-generative-ai-on-productivity-innovation-and-entrepreneurship_da1d085d/b21df222-en.pdf?utm_source=chatgpt.com

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., ... and Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS nexus, 3(6), page 191.

Chakravorti, Tatiana, Xinyu Wang, Pranav Narayanan Venkit, Sai Koneru, Kevin Munger, and Sarah Rajtmajer. 2025. “Social Scientists on the Role of AI in Research.” Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 8(1). https://doi.org/10.1609/aies.v8i1.36568

Challapally, A., Pease, C., Raskar, R., and Chari, P. (2025). The GenAI divide: State of AI in business 2025. Artificial Intelligence News. https://www.artificialintelligence-news.com/wp-content/uploads/2025/08/ai_report_2025.pdf

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., and Wadman, K. (2025). How People Use ChatGPT (No. w34255). National Bureau of Economic Research.

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., and Zemmel, R. (2023). *The economic potential of generative AI: The next productivity frontier*. McKinsey Global Institute. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Cottier, B., Snodin, B., Owen, D., and Adamczewski, T. (2025, March 12). LLM inference prices have fallen rapidly but unequally across tasks. Epoch AI. https://epoch.ai/data-insights/llm-inference-price-trends

Doshi, A. R., and Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

DeVerna, M. R., Yan, H. Y., Yang, K.-C., and Menczer, F. (2024). *Fact-checking information from large language models can decrease headline discernment*. Proceedings of the National Academy of Sciences, 121(50), e2322823121. https://doi.org/10.1073/pnas.2322823121

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130, 10; https://arxiv.org/abs/2303.10130

Henshall, Will. "The Billion-Dollar Price Tag of Building AI." TIME, 3 June 2024, https://time.com/ai-costing-more-and-more-to-train.

Hoffmann, M., Boysel, S., Nagle, F., Peng, S., and Xu, K. (2024). Generative AI and distributed work: Evidence from open source software. Unpublished. Version: September, 4, 2024. https://thedocs.worldbank.org/en/doc/d09a6806e7d7efb816af153002261f1e-0070012021/related/Hoffmann-Copilot-and-Distributed-Work.pdf

Humlum, A., and Vestergaard, E. (2024). The unequal adoption of ChatGPT exacerbates existing inequalities among workers. Proceedings of the National Academy of Sciences of the United States of America, 122(1), e2414972121. https://doi.org/10.1073/pnas.2414972121

Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., and Dean, R. (2025). AI 2027. AI Futures Project. https://ai-2027.com/

Korinek, A., and Suh, D. (2024). Scenarios for the Transition to AGI (No. w32255). National Bureau of Economic Research.

Kumar, H., Rothschild, D. M., Goldstein, D. G., and Hofman, J. M. (2023). Math education with large language models: Peril or promise?. Available at SSRN 4641653.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. (2024). *Hallucination-free? Assessing the reliability of leading AI legal research tools* (Preprint).

Marguerit, D. (2025). Augmenting or automating labor? The effect of AI development on new work, employment, and wages (arXiv:2503.19159). arXiv. https://doi.org/10.48550/arXiv.2503.19159

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... and Mian, A. (2025). A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology, 16(5), 1-72. Accessed on October 31, 2025 from: https://arxiv.org/pdf/2307.06435.

Niederhoffer, K., Kellerman, G. R., Lee, A., Liebscher, A., Rapuano, K., and Hancock, J. T. (2025). AI-generated "workslop" is destroying productivity. Harvard Business Review.

Rothschild, D. M., Mobius, M., Hofman, J. M., Dillon, E. W., Goldstein, D. G., Immorlica, N., ... and Vogel, M. (2025). The Agentic Economy. arXiv preprint arXiv:2505.15799

Simchon, A., Edwards, M., and Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative artificial intelligence. PNAS nexus, 3(2), pgae035.

Spitale, G., Biller-Andorno, N., and Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. Science Advances, 9(26), eadh1850.

Tomlinson, K., Jaffe, S., Wang, W., Counts, S., and Suri, S. (2025). Working with AI: Measuring the applicability of generative AI to occupations (arXiv:2507.07935). arXiv. https://arxiv.org/abs/2507.07935

Washington Post Editorial Board. (2025). AI is reshaping jobs - but here's what companies aren't saying about layoffs [Interactive feature]. The Washington Post. https://www.washingtonpost.com/opinions/interactive/2025/ai-jobs-layoffs-tech/

Zavolokina, Liudmila, et al. "Think fast, think slow, think critical: designing an automated propaganda detection tool." Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024.

Zhou, E., and Lee, D. (2024). Generative artificial intelligence, human creativity, and art. PNAS Nexus, 3(3), pgae052. https://doi.org/10.1093/pnasnexus/pgae052

Section 3

Adhikari, D. M., Hartland, A., Weber, I., and Cannanure, V. K. (2025, July). Exploring LLMs for automated generation and adaptation of questionnaires. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces* (pp. 1-23).

Allamong, M. B., Jeong, J., and Kellstedt, P. M. (2025). Spelling correction with large language models to reduce measurement error in open-ended survey responses. Research and Politics, 12(1), 20531680241311510.

Anthis, Jacy Reese, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. 2025. "Position: LLM Social Simulations Are a Promising Research Method." In Forty-second International Conference on Machine Learning Position Paper Track.

Argyle, Lisa P., Busby, Ethan C., Gubler, Joshua R., Hepner, Bryce, Lyman, Alex, and Wingate David. 2025. "Arti-'fickle' intelligence: using LLMs as a tool for inference in the political and social sciences". Nature Computational Science 5, 737–744. https://doi.org/10.1038/s43588-025-00843-4

Barari, S., Angbazo, J., Wang, N., Christian, L. M., Dean, E., Slowinski, Z., and Sepulvado, B. (2025). AI-assisted conversational interviewing: Effects on data quality and respondent experience. arXiv. https://doi.org/10.48550/arXiv.2504.13908

Barends, A. J., and de Vries, R. E. (2025). Developing and improving personality inventories using generative artificial intelligence: The psychometric properties of a short HEXACO scale developed using ChatGPT 4.0. *Journal of Personality Assessment*, *107*(4), 419-425.

Behrend, T. S., and Landers, R. N. (2025). Participant Interactions with Artificial Intelligence: Using Large Language Models to Generate Research Materials for Surveys and Experiments. Journal of Business and Psychology, 1-23.

Beltoft, S. L., Schneider-Kamp, P., and Askegaard, S. T. (2025). Interview Bot: Can Agentic LLM's Perform Ethnographic Interviews?. In 17th International Conference on Agents and Artificial Intelligence, ICAART 2025 (pp. 702-709).

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." Political Analysis 32, no. 4: 401–16.

Bryda, G., and Sadowski, D. (2024). From words to themes: AI-powered qualitative data coding and analysis. In Computer Supported Qualitative Research (pp. 309–345). Springer. https://doi.org/10.1007/978-3-031-65735-1_19

Buskirk, T. D., Keusch, F., von der Heyde, L., and Eck, A. (2025a). More Parameters Than Populations: A Systematic Literature Review of Large Language Models within Survey Research. arXiv preprint arXiv:2509.03391. https://arxiv.org/abs/2509.03391

Buskirk, T.D., Eck, A., and Timbrook, J. (2025b). The Task Is to Improve the Ask: An Experiment for Developing Prompts to Generate High Quality Survey Items from Large Language Models. Available at: http://dx.doi.org/10.2139/ssrn.5377878

Buskirk, T.D., Eck, A., Timbrook, J. and Tatum, H. (2025c) Is Your Chatbot Smarter Than a 5th Grader? an Experiment Testing the Steerability of Reading Levels of Survey Questions Created Using Generative AI Tools. Paper presented at the 80th Annual American Association of Public Opinion Research Conference, Saint Louis, MO May 14-16, 2025. https://aapor.confex.com/aapor/2025/meetingapp.cgi/Paper/3980

Buskirk, T.D., Lerner, J. and Benson, A. (2025d) What Happens When We Prompt the Model, Not the People: LLMs as Survey Estimators. Paper presented at the 50th Annual Midwest Association of Public Opinion Research Conference, Chicago, IL, November 21-22, 2025.

Côté, P.-O., Nikanjam, A., Ahmed, N., Humeniuk, D., and Khomh, F. (2024). *Data cleaning and machine learning: A systematic literature review*. Automated Software Engineering, 31(2), 54. https://doi.org/10.1007/s10515-024-00453-w

Cuevas, A., Scurrell, J. V., Brown, E. M., Entenmann, J., and Daepp, M. I. (2025). Collecting Qualitative Data at Scale with Large Language Models: A Case Study. Proceedings of the ACM on Human-Computer Interaction, 9(2), 1-27. Available from: https://dl.acm.org/doi/10.1145/3710947

Ehrett C, Hegde S, Andre K, Liu D, Wilson T. Leveraging Open-Source Large Language Models for Data Augmentation in Hospital Staff Surveys: Mixed Methods Study. JMIR Med Educ. 2024 Nov 19;10:e51433. doi: 10.2196/51433. PMID: 39560937; PMCID: PMC11590755.

Eftekhari, H. (2024). Transcribing in the digital age: Qualitative research practice utilizing intelligent speech recognition technology. European Journal of Cardiovascular Nursing, 23(5), 553–560. https://doi.org/10.1093/eurjcn/zvae013 [academic.oup.com]

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130, 10; https://arxiv.org/abs/2303.10130

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

Gweon, H., and Schonlau, M. (2024). *Automated classification for open-ended questions with BERT*. Journal of Survey Statistics and Methodology, 12(2), 493–504. https://doi.org/10.1093/jssam/smad015

Hernandez, I., and Nie, W. (2023). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. Personnel Psychology, 76(4), 1011-1035.

Ilyas, I. F., and Rekatsinas, T. (2022). Machine learning and data cleaning: Which serves the other?. *ACM Journal of Data and Information Quality (JDIQ)*, *14*(3), 1-11.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., and Gimenez, A. (2016). Evaluating online nonprobability surveys. Pew Research Center, 61, 1-60. Available at: https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/

Kuo, S.-M., Tai, S.-K., Lin, H.-Y., and Chen, R.-C. (2025). Automated clinical trial data analysis and report generation by integrating retrieval-augmented generation and large language models. AI, 6(8), 188. https://doi.org/10.3390/ai6080188

Lee, Sunghee, Tian, J. and Morales, S. (2025). Evaluation of AI-Assisted Survey Questionnaire Translation. Paper presented at the 50th Annual Midwest Association of Public Opinion Research Conference, November 20-21, 2025.

Mellon, Jonathan, Jack Bailey, Rosie Scott, Johannes Breckwoldt, Marco Miori, and Peter Schmedeman. 2024. “Do AIs Know What the Most Important Issue Is? Using Language Models to Code Open-Text Social Survey Responses at Scale.” Research and Politics 11, no. 1. https://doi.org/10.1177/20531680241231468.

Metheney, E. A., and Yehle, L. (2024). Exploring the Potential Role of Generative AI in the TRAPD Procedure for Survey Translation. *arXiv preprint arXiv:2411.14472*.

Mojadeddi, Z. M., and Rosenberg, J. (2024). Automated transcription of interviews in qualitative research using artificial intelligence: A simple guide. Journal of Surgery Research and Practice, 5(2), 1–6. https://doi.org/10.46889/JSRP.2024/5204 [research.regionh.dk]

Moon, S., Abdulhai, M., Kang, M., Suh, J., Soedarmadji, W., Behar, E. K., and Chan, D. M. (2024). Virtual personas for language models via an anthology of backstories. arXiv preprint arXiv:2407.06576; Available at: https://arxiv.org/abs/2407.06576

Morris, G. Elliott, Benjamin Leff, and Peter K. Enns. 2025. “The Limits of Synthetic Samples in Survey Research” Verasight White Paper Series. https://www.verasight.io/reports/synthetic-sampling-2

Padgett, Z., et al. 2024. “Evaluating the Quality of Questionnaires Created with SurveyMonkey’s Build with AI.” Paper presented at the 79th Annual Conference of the American Association for Public Opinion Research, Atlanta, GA, May 15–17. Accessed June 20, 2024. https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/3198.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. “Generative Agent Simulations of 1,000 People.” arXiv preprint arXiv:2411.10109. https://doi.org/10.48550/arXiv.2411.10109.

Rogers, R., and Zhang, X. (2024). The Russia–Ukraine War in Chinese Social Media: LLM Analysis Yields a Bias Toward Neutrality. *Social Media + Society*, *10*(2).

Rothschild, D. M., Buskirk, T. D., Eckman, S., Hillygus, D. S., Kreuter, F., and Lazer, D. (2025). Successfully navigating the disruption AI will bring to survey research. The Survey Statistician, 92, 30–44. Retrieved from https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2025_July_N92_04.pdf

Rothschild, David., The Economics of Transformative AI (University of Chicago Press, 2025), https://www.nber.org/books-and-chapters/economics-transformative-ai/comment-coasean-singularity-demand-supply-and-market-design-ai-agents-rothschild.

Rothschild, D. M., Brand, J., Schroeder, H., and Wang, J. (2024). Opportunities and risks of LLMs in survey research. Available at SSRN.

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., and Wingate, D. (2023). Towards coding social science datasets with language models. arXiv preprint arXiv:2306.02177; Available at: https://arxiv.org/abs/2306.02177

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023, July). Whose opinions do language models reflect? In *International conference on machine learning* (pp. 29971-30004). PMLR.

Shedlock, Ashley, (2025) "Smarter Surveys, Faster: How AI is Transforming Survey Design", https://www.greenbook.org/insights/the-prompt-ai/smarter-surveys-faster-how-ai-is-transforming-survey-design

Soliman, H., and Gurevych, I. (2025). A survey on advances in retrieval-augmented generation over tabular data and table QA. OpenReview. https://openreview.net/forum?id=AdDU2c4XfP

Tirumala, S., Jain, N., Leybzon, D. D., and Buskirk, T. D. (2025). Mic Drop or Data Flop? Evaluating the Fitness for Purpose of AI Voice Interviewers for Data Collection within Quantitative and Qualitative Research Contexts. Presented and published as a conference paper at COLM 2025 NLPOR Workshop, Montreal, Canada, October 10, 2026. Available at: https://openreview.net/pdf?id=Z4vRAcchxt

Tewari, T., and Hosein, P. (2024). Automating the Conducting of Surveys Using Large Language Models. In *International Conference on Deep Learning Theory and Applications* (pp. 136-151). Cham: Springer Nature Switzerland.

von der Heyde, L., Haensch, A. C., and Wenz, A. (2025a). Vox populi, vox ai? using large language models to estimate german vote choice. Social Science Computer Review, 08944393251337014.

Von Der Heyde, L., Haensch, A. C., Weiß, B., and Daikeler, J. (2025, December). Using large language models for coding german open-ended survey responses on survey motivation. In Survey Research Methods (Vol. 19, No. 4, pp. 355-370).

Wang, A., Morgenstern, J., and Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. Nature Machine Intelligence, 7(3), 400-411.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. International journal of forecasting, 31(3), 980-991.

Wuttke, A., Aßenmacher, M., Klamm, C., Lang, M., and Kreuter, F. (2025, May). AI conversational interviewing: Transforming surveys with LLMs as adaptive interviewers. In Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025) (pp. 179-204).

Xu, W., Mao, Y., Zhang, X., Zhang, C., Dong, X., and Gao, Y. (2025). DAgent: A relational database-driven data analysis report generation agent. arXiv. https://doi.org/10.48550/arXiv.2503.13269

Yun, H. S., Arjmand, M., Sherlock, P., Paasche-Orlow, M. K., Griffith, J. W., and Bickmore, T. (2023). Keeping users engaged during repeated administration of the same questionnaire: Using large language models to reliably diversify questions. arXiv preprint arXiv:2311.12707; https://arxiv.org/html/2311.12707v2

Section 4

Alansari, A., and Luqman, H. (2025). Large language models hallucination: A comprehensive survey. *arXiv preprint arXiv:2510.06265*. https://arxiv.org/abs/2510.06265

Amaya, A., Biemer, P. P., and Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology*, 8(1), 89-119. https://academic.oup.com/jssam/article-abstract/8/1/89/5728725

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669-674. https://www.science.org/doi/10.1126/science.adi6000

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337-351. https://doi.org/10.1017/pan.2023.2

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

Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., ... and Mahdi, A. (2025). Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*. https://arxiv.org/abs/2511.04703

Broska, D., Howes, M., and van Loon, A. (2025). The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations. *Sociological Methods and Research*, 00491241251326865. https://doi.org/10.1177/00491241251326865

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... and Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901. https://arxiv.org/abs/2005.14165

Chen, Y., Yin, C.H., Chikodikar, S.M. and Vinayak, R.K., On the Scoring Functions for RAG-based Conformal Factuality. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Chen, Y., Chen, D., Chikodikar, S. M., Yin, C. H., and Vinayak, R. K. (2026). Is Conformal Factuality for RAG-based LLMs Robust? Novel Metrics and Systematic Insights. *arXiv preprint arXiv:2603.16817*.

Conrad, F. G., and Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, *64*(1), 1-28. https://doi.org/10.1086/316757

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

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: Wiley.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... and Guo, J. (2024). A survey on llm-as-a-judge. *The Innovation*. https://arxiv.org/abs/2411.15594

Guerreiro, N. M., Voita, E., and Martins, A. F. (2023, May). Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1059-1075). https://aclanthology.org/2023.eacl-main.75.pdf

Hakim, J. B., Painter, J. L., Ramcharran, D., Kara, V., Powell, G., Sobczak, P., ... and Beam, A. (2025). The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings. *Scientific Reports*, *15*(1), 27886. https://www.nature.com/articles/s41598-025-09138-0

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. *Springer Series in Statistics Springer.* https://doi.org/10.1007/978 -0-387-84858-7

He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S., and Wang, X. (2024). *Exploring human-like translation strategy with large language models*. Transactions of the Association for Computational Linguistics, 12, 229–246. https://doi.org/10.1162/tacl_a_00642

Hullman, J., Broska, D., Sun, H., and Shaw, A. (2025). This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. https://mucollective.northwestern.edu/files/Hullman-llm-behavioral.pdf

Krsteski, S., Russo, G., Chang, S., West, R., and Gligorić, K. (2025). Valid survey simulations with limited human data: The roles of prompting, fine-tuning, and rectification. *arXiv preprint arXiv:2510.11408*. https://arxiv.org/abs/2510.11408

Kuha, J., Butt, S., Katsikatsou, M., and Skinner, C. J. (2018). The effect of probing "don't know" responses on measurement quality and nonresponse in surveys. *Journal of the American Statistical Association*, *113*(521), 26-40. https://doi.org/10.1080/01621459.2017.1323640

Laurito, W., Davis, B., Grietzer, P., Gavenčiak, T., Böhm, A., and Kulveit, J. (2025). AI–AI bias: Large language models favor communications generated by large language models. *Proceedings of the National Academy of Sciences*, *122*(31), e2415697122. https://www.pnas.org/doi/10.1073/pnas.2415697122

Lee, Sunghee, Tian, J. and Morales, S. (2025). Evaluation of AI-Assisted Survey Questionnaire Translation. Paper presented at the 50th Annual Midwest Association of Public Opinion Research Conference, November 20-21, 2025.

Lyman, Alex, et al. "Balancing large language model alignment and algorithmic fidelity in social science research." *Sociological Methods and Research* 54.3 (2025): 1110-1155. https://doi.org/10.1177/00491241251342008

Mellon, Jonathan, Jack Bailey, Rosie Scott, Johannes Breckwoldt, Marco Miori, and Peter Schmedeman. 2024. "Do AIs Know What the Most Important Issue Is? Using Language Models to Code Open-Text Social Survey Responses at Scale." Research and Politics 11, no. 1. https://doi.org/10.1177/20531680241231468.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). https://dl.acm.org/doi/pdf/10.1145/3287560.3287596

Naidu, Gireen, Tranos Zuva, and Elias Mmbongeni Sibanda. "A review of evaluation metrics in machine learning algorithms." *Computer science on-line conference*. Cham: Springer International Publishing, 2023. https://doi.org/10.1007/978-3-031-35314-7_2

National Institute of Standards and Technology (NIST) (2023). *Glossary*. NIST Trustworthy and Responsible AI Resource Center. https://airc.nist.gov/glossary/

NIST (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile. *NIST Trustworthy and Responsible AI Gaithersburg, MD, USA*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

Palmer, A., Smith, N. A., and Spirling, A. (2024). Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, *4*(1), 2-3. https://www.nature.com/articles/s43588-023-00585-1

Rabanser, Stephan, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. "Towards a Science of AI Agent Reliability." *arXiv preprint arXiv:2602.16666* (2026). https://arxiv.org/pdf/2602.16666v1

Rao, Anita, Drew Keller, Neha Kalra, Ryan Steed, Kweku Kwegyir-Aggrey, Kevin Klyman, Diane Staheli, and Stevie Bergman. "Challenges to the Monitoring of Deployed AI Systems." (2026). *NIST Trustworthy and Responsible AI*, NIST AI 800-4. https://doi.org/10.6028/NIST.AI.800-4

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2024). *Quantifying language models' sensitivity to spurious features in prompt design (or: How I learned to start worrying about prompt formatting)*. International Conference on Learning Representations (ICLR 2024). https://arxiv.org/abs/2310.11324

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E.H., Schärli, N. and Zhou, D., 2023, July. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning* (pp. 31210-31227). PMLR.

Sivaprasad, S., Kaushik, P., Abdelnabi, S., and Fritz, M. (2025, July). A theory of response sampling in LLMs: Part descriptive and part prescriptive. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 30091-30135).

Statistics Canada. (2017). *Statistics Canada's quality assurance framework* (3rd ed.). Ottawa, ON: Statistics Canada.

Tabassi, E., and National Institute of Standards and Technology. (2023, January 26). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Vishwakarma, H., Lin, H., Sala, F. and Korlakai Vinayak, R., 2023. Promises and pitfalls of threshold-based auto-labeling. *Advances in Neural Information Processing Systems*, *36*, pp.51955-51990. https://proceedings.neurips.cc/paper_files/paper/2023/file/a355051cc32d36e2a971de190701745a-Paper-Conference.pdf

Vishwakarma, H., Lin, H. and Vinayak, R.K., 2024, April. Taming False Positives in Out-of-Distribution Detection with Human Feedback. In *International Conference on Artificial Intelligence and Statistics* (pp. 1486-1494). PMLR.

Von Der Heyde, L., Haensch, A. C., Weiß, B., and Daikeler, J. (2025, December). Using large language models for coding german open-ended survey responses on survey motivation. In Survey Research Methods (Vol. 19, No. 4, pp. 355-370).

Tonneau, M., Seghal, N. K., Malhotra, N., Orozco-Olvera, V., Boudet, A. M. M., Subramanian, L., ... and Hofmann, V. (2026). Demographic Probing of Large Language Models Lacks Construct Validity. *arXiv preprint arXiv:2601.18486*. https://arxiv.org/abs/2601.18486

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., and Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, 595(7866), 197-204. https://www.nature.com/articles/s41586-021-03666-1

Weeber, F., Neplenbroek, V., Batzner, J., and Padó, S. (2026). One Persona, Many Cues, Different Results: How Sociodemographic Cues Impact LLM Personalization. *arXiv preprint arXiv:2601.18572*. https://arxiv.org/abs/2601.18572

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* NeurIPS. https://dl.acm.org/doi/10.5555/3600270.3602070

West, B. T., Conrad, F. G., Kreuter, F., and Mittereder, F. (2018). Can conversational interviewing improve survey response quality without increasing interviewer effects?. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(1), 181-203. https://doi.org/10.1111/rssa.12255

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1-9. https://www.nature.com/articles/sdata201618

Yun, H. S., Arjmand, M., Sherlock, P., Paasche-Orlow, M. K., Griffith, J. W., and Bickmore, T. (2023). Keeping users engaged during repeated administration of the same questionnaire: Using large language models to reliably diversify questions. *arXiv preprint arXiv:2311.12707*. https://arxiv.org/abs/2311.12707

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*. https://arxiv.org/abs/2307.15043

Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, *36*, 46595-46623. https://neurips.cc/virtual/2023/poster/73434

Section 5

Abraham L, Arnal C, and Marie A (2025) Prompt selection matters: enhancing text annotations for social sciences with large language models, https://doi.org/10.1007/s42001-025-00388-6.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate (2023) Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31 (2023): 337-351. https://doi.org/10.1017/pan.2023.2.

Bender, E. M., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *https://dl.acm.org/doi/10.1145/3442188.3445922*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229).
https://dl.acm.org/doi/pdf/10.1145/3287560.3287596

Suh, A., Hurley, I., Smith, N., and Siu, H. C. (2025). Fewer than 1% of explainable AI papers validate explainability with humans. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, April 26–May 1, 2025, Yokohama, Japan.
https://doi.org/10.1145/3706599.3719964

Wan, A., Klyman, K., Kapoor, S., Maslej, N., Longpre, S., Xiong, B., Liang, P., Bommasani, R. (2025). The 2025 Foundational Model Transparency Index.
https://crfm.stanford.edu/fmti/December-2025/paper.pdf

Section 6

Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Dass, R. K., Petersen, N., Omori, M., Lave, T. R., and Visser, U. (2023). Detecting racial inequalities in criminal justice: towards an equitable deep learning approach for generating and interpreting racial categories using mugshots. AI and SOCIETY, 38(2), 897-918.

Federal Trade Commission. (2022). Combatting online harms through innovation.
https://www.ftc.gov/reports/combatting-online-harms-through-innovation

Gray, M. L., and Suri, S. (2019). Ghost work: How to stop Silicon Valley from building a new global underclass. Harper Business.

Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., ... and Rashidi, H. H. (2025). Ethical and bias considerations in artificial intelligence/machine learning. Modern Pathology, 38(3), 100686.

Hao, K. (2025). Empire of AI: Dreams and nightmares in Sam Altman's OpenAI. Penguin Group.

Hartung, T., and Kleinstreuer, N. (2025). Challenges and opportunities for validation of AI-based new approach methods. ALTEX, 42(1), 3–21. https://doi.org/10.14573/altex.2412291

Kim, J. H., Kim, J., Park, J., Kim, C., Jhang, J., and King, B. (2025). When ChatGPT gives incorrect answers: the impact of inaccurate information by generative AI on tourism decision-making. Journal of Travel Research, 64(1), 51-73.

Kshetri, N. (2024). Linguistic challenges in generative artificial intelligence: Implications for low-resource languages in the developing world. Journal of Global Information Technology Management, 27(2), 95-99.

Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. Race and class, 60(4), 3-26.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research. U.S. Department of Health, Education, and Welfare. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., and Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. Patterns, 2(10).

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Oladokun, B., Emmanuel, V., Osagie, O. Q., and Alabi, A. (2025). *ChatGPT and library users: AI risks of hallucinations and misinformation*. Cybrarians Journal, 76. https://doi.org/10.70000/cj.2025.76.642

Pew Research Center (2025). How People Around the World View AI. Retrieved from https://www.pewresearch.org/wp-content/uploads/sites/20/2025/10/pg_2025.10.15_ai_report.pdf

Sakai, Y., Kamigaito, H., and Watanabe, T. (2026). *HalluCitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in ACL conferences*. arXiv:2601.18724. https://arxiv.org/abs/2601.18724

Salganik, M. J. (2019). Bit by bit: Social research in the digital age. Princeton University Press.

Secretary's Advisory Committee on Human Research Protections (2022). IRB Considerations on the Use of Artificial Intelligence in Human Subjects Research. US Department of Health and Human Services. Retreived from https://www.hhs.gov/ohrp/sachrp-

committee/recommendations/irb-considerations-use-artificial-intelligence-human-subjects-research/index.html

Sun, Y., Sheng, D., Zhou, Z., and Wu, Y. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications, 11(1), 1-14.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems, 10(05), 557-570.

Tacheva, J., and Ramasubramanian, S. (2023). AI Empire: Unraveling the interlocking systems of oppression in generative AI's global order. Big Data and Society, 10(2), 20539517231219241.

United Nations Educational, Scientific and Cultural Organization (2022). Recommendation on the Ethics of Artificial Intelligence. Retrieved from https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

US Department of Homeland Security (2012). The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. Retrieved from https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf

Wiegand, T. L. T., Jung, L. B., Gudera, J. A., Schuhmacher, L. S., Moehrle, P., Rischewski, J. F., ... and Koerte, I. K. (2025). Demographic inaccuracies and biases in the depiction of patients by artificial intelligence text-to-image generators. NPJ Digital Medicine, 8(1), 459.

Zaretsky, J., Kim, J. M., Baskharoun, S., Zhao, Y., Austrian, J., Aphinyanaphongs, Y., ... and Feldman, J. (2024). Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA network open, 7(3), e240357-e240357.