

Let's Not Leave
Probability Panels to **Chance!**
Why AI Matters for Their Future

**Designing Smarter, More Resilient
Probability Panels for the AI Era**

Trent D. Buskirk

Professor and Provost Fellow

School of Data Science and Joint School of Public Health

Old Dominion University

CIPHER Keynote 2026

Field Methods

Volume 34, Issue 1, February 2022, Pages 20-35

© The Author(s) 2022, Article Reuse Guidelines

<https://doi-org.proxy.lib.ou.edu/10.1177/1525822X211069640>



Article

Scratch the Scratch-off: Testing Prepaid and Conditional Incentives with Postcard and Letter Invitations in a Web-push Design with an Address-based Sample

Philip S. Brenner¹ and Trent D. Buskirk²

Abstract

We tested a novel extension to mailed invitations to a web-push survey, using a postcard invitation to deliver a scratch-off giftcode incentive similar to an instant-win lottery ticket. Scratch-off postcards were included as one of five conditions in randomized survey experiment varying two mailing types (letter and postcard) and three incentive types (prepaid cash, prepaid giftcodes, and conditional giftcodes). Invitations were sent to a sample of 17,808 addresses in Boston, Massachusetts, recruiting for a new online panel study of city residents. We report response rates and costs for each condition.



BEACON
BE HEARD BOSTON!

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

The road ahead...

- My experience with the Beacon panel isn't unlike the experiences of others who develop, use and maintain probability panels. We all know that:

**Probability panels are built on chance,
but their future shouldn't be left to it.**

**The future of probability panels won't be
determined by how random our samples are—
but by how intentional our designs become.**

The road ahead is NOW!

“Probability sampling is NOT dead yet—not by far. But it will starve to death if no one invests in its modernization...

We must protect what we know works, while of course remaining unafraid to experiment...”

– Torbjörn Sjöström, AAPORnet Post, February 18, 2026

Thinking about the Future...

- The scientific advantages of probability panels remain clear.
 - I know I am preaching to the proverbial choir here 😊
- But the operational realities of sustaining them are becoming more challenging —
 - participation declines
 - costs increasing
 - Increased competition (alternate data sources and nonprobability panels)
 - Continued expectations for timely, high-quality, but cost effective data
- And this raises a design question for the AI era: how do we use new tools to refine and sustain probability panels, in particular, to protect their use moving forward —
 - not by replacing chance, but by designing more intelligently around it?

Refining our Design?

- In AI, we often talk about keeping humans in the loop — as if human involvement is something that needs to be preserved.
- But probability panels have always been human-in-the-loop systems.
 - They rely on people not only as participants but also expertise from humans for design, analysis, inference and interpretation.
- The design question in my mind isn't whether humans are in the loop — they always are.
- The real question is where does AI belongs in this loop?

Bringing AI into our Design?

- Used thoughtfully, AI can enable optimization in several parts of the loop that are operationally complex, data-rich, and time-sensitive —
 - helping us detect problems earlier, adapt more intelligently, and reduce burden on both respondents and panel infrastructure.
- But there are parts of the loop where AI should inform decisions, not make them — particularly where inference, transparency, reproducibility and scientific validity are concerned.
- Designing smarter probability panels in the AI era requires being clear about this distinction: using AI to improve how panels operate, while ensuring that inference and scientific validity remain grounded in human design and statistical reasoning.



The LAIandscape Forward...

- Smarter probability panels won't necessarily be fully automated — but current trends suggest they will become intentionally augmented with AI thoughtfully placed in the loop.
- AI may not operate within just one part of the loop, but could function independently and in tandem in several parts of it.
 - Either as AI enabled software/service-ware or as one or more Agents, for example.



Tracing AI Use in Surveys...

- Barari et al. (2025) and Rothchild et al. (2025) have proposed the framing of Role-Based considerations for AI in Survey Research:
 - AI as Assistant/Research Colleague
 - AI as the Interviewer
 - AI as the Respondent
 - AI as Processor/Labeler/Modeler
 - AI as Briefer/Reporter
- Buskirk and colleagues (2025) report on the early findings from a systematic literature review that frames the current work in using LLMs within Survey Research using a process and task-specific lens – LLMs being deployed for tasks within:
 - Pre-data Collection Phase (e.g. questionnaire and sample design; pretesting, etc.)
 - Data Collection Phase (e.g. data retrieval; synthetic response generation, etc.)
 - Post-Data Collection Phase (e.g. data processing/coding, analysis/modeling, etc.)

The LAIandscape Forward...

- As AI enters probability panels in these different roles—as interviewer, respondent, designer, and analyst—the question shifts from whether it can be used to whether it should be used, and how.
- Working in this space over the past 5 years I have found it useful to think about this in terms of both promise and precaution.
 - At each point in the panel workflow, AI introduces new capabilities—but it also changes how data are generated, how decisions are made, and how information is processed.
 - These changes can amplify known sources of error and introduce new ones that are not yet fully understood.
- The literature is beginning to provide concrete examples of both—applications that show real promise, and others that highlight the need for precaution.



Promises and Precautions of AI in Survey Research

- With a focal point of probability-based panels and framing this work through the lens of Promises and Precautions, I will highlight :
 - **Where AI has already been introduced into the panel workflow**
 - **What has shown promise—and what has not**
 - **What precautions have we already identified**
 - **What new risks and threats this technology may introduce**
 - **And how probability panels themselves can play a critical role in shaping and strengthening the emerging AI ecosystem**
 - **Ways to leverage historical Panel Data Assets to Improve Operations**



So what are LLM's Anyway?

They learn patterns and relationships from large volumes of textual data to understand the structure of a language. These patterns are encoded into millions and billions of parameters that create the “model.”

These models are ***not deterministic*** so they can generate diverse and creative responses.

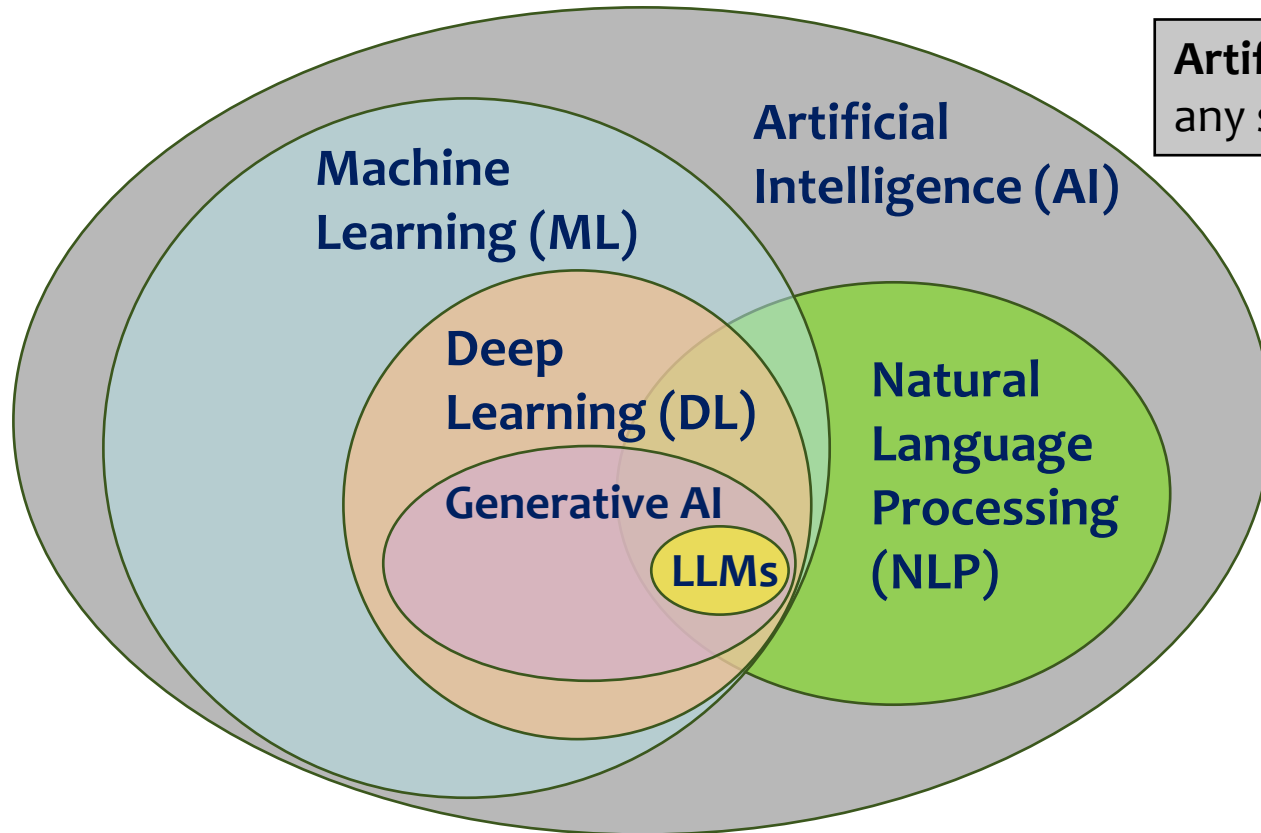
LLMs

These models can then be used to generate new text based on inputs by predicting the most probable sequence of words to follow...

A semi-technical, detailed and comprehensive overview of LLMs: <https://arxiv.org/pdf/2307.06435>

The AI Hierarchy

Source Attribution: [Gemini](#) based on prompt: “can you generate a diagram that shows the relationship between ML, AI, Deep Learning, NLP, Generative AI and LLMs”, October 2024.



Artificial Intelligence (AI): The broadest concept, encompassing any system that can exhibit human-like intelligence.

Machine Learning (ML): A subfield of AI that focuses on algorithms that can learn from data without explicit programming.

Deep Learning: A subfield of ML inspired by the structure and function of the brain. It uses artificial neural networks with multiple layers to learn complex patterns from large amounts of data.

Generative AI: A subfield of AI focused on algorithms that can create new content, like text, images, or music.

Large Language Models (LLMs): A type of generative AI model trained on massive amounts of text data to create human-quality text in a variety of applications.

Natural Language Processing (NLP): A subfield of AI concerned with the interaction between computers and human language. NLP tasks include text generation, translation, sentiment analysis, and question answering.

Tracing LLM Use within Survey Research

- **Buskirk and colleagues (2025) report the preliminary findings from a systematic literature review of LLMs being used to complete tasks within the survey research process.**
 - Queried 3 Databases using 62 “Survey” and 10 “LLM” keywords that appeared between 2019 and 2025:
 - Semantic Scholar
 - Web of Science
 - arXiv
 - To be included in the final review paper had to use **LLMs within the survey research process**
 - Data could be survey specific or an alternate data source being used for generating/understanding public opinion, broadly defined
 - In total we had 136 eligible papers for final, full review

More Parameters Than Populations: A Systematic Literature Review of Large Language Models within Survey Research

Trent D Buskirk
School of Data Science
Old Dominion University
Norfolk, VA 23529, USA
tbuskirk@odu.edu

Florian Keusch
School of Social Sciences
University of Mannheim
Mannheim, Germany
f.keusch@uni-mannheim.de

Leah von der Heyde
Department of Statistics
LMU Munich
Munich, Germany
l.heyde@lmu.de

Adam Eck
Computer Science Department
Oberlin College
Oberlin, OH 44074, USA
aeck@oberlin.edu

Abstract

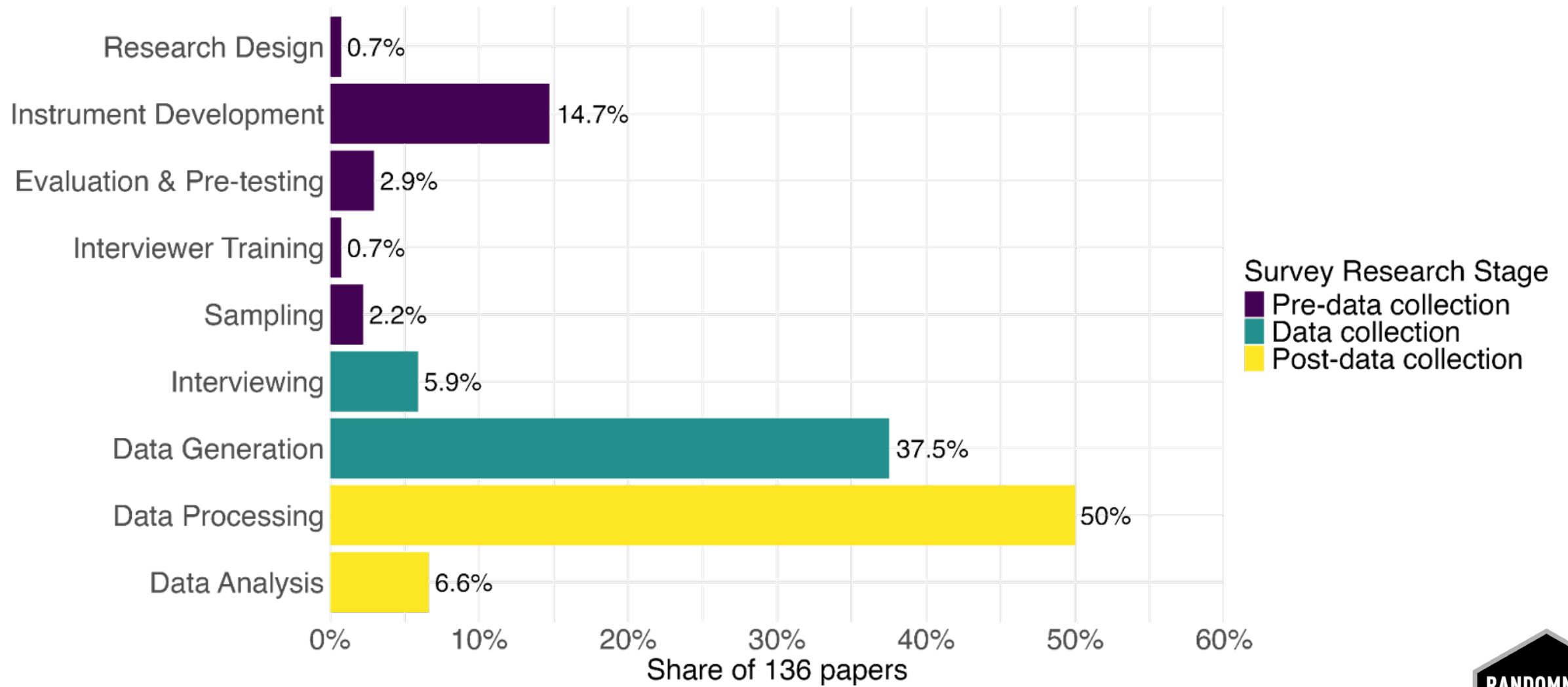
Survey research has a long-standing history of being a human-powered field, but one that embraces various technologies for the collection, processing, and analysis of various behavioral, political, and social outcomes of interest, among others. At the same time, Large Language Models (LLMs) bring new technological challenges and prerequisites in order to fully harness their potential. In this paper, we report work-in-progress on a systematic literature review based on keyword searches from multiple

[Preliminary Work:](https://openreview.net/pdf?id=0Hxhwa56Yg)

<https://openreview.net/pdf?id=0Hxhwa56Yg>

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

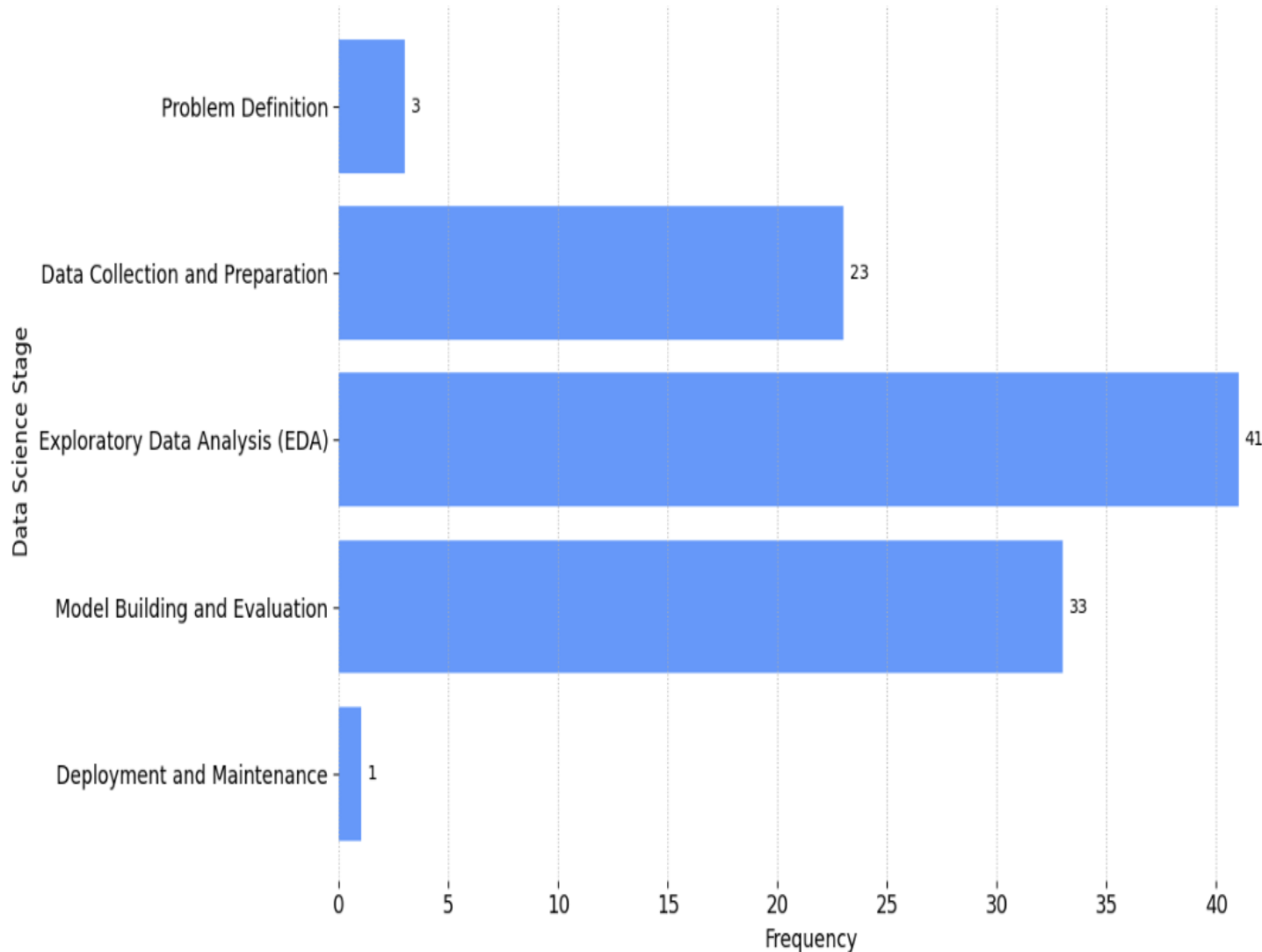
Tracing LLM Use in Survey Research...



Keusch et al. (2026, forthcoming work).

Preliminary work has appeared here: <https://openreview.net/pdf?id=0Hxhwa56Yg>

Tracing LLM Use Data Science



Large Language Models in the Data Science Lifecycle: A Systematic Mapping Study

Sai Sanjna Chintakunta¹, Nathalia Nascimento^{1†}, Everton Guimaraes^{1†}
Engineering Division, Great Valley, The Pennsylvania State University, Pennsylvania,
United States.

Contributing authors: sqc6557@psu.edu; nqm5742@psu.edu; ezt157@psu.edu;
[†]These authors contributed equally to this work.

Abstract

In recent years, Large Language Models (LLMs) have emerged as transformative tools across numerous domains, impacting how professionals approach complex analytical tasks. This systematic mapping study comprehensively examines the application of LLMs throughout the Data Science lifecycle. By

Attribution: <https://arxiv.org/pdf/2508.11698>

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

The Promise of LLMs within the Pre-Data Collection Phase: AI as Collaborator/Assistant



In their article "Keeping Users Engaged During Repeated Administration of the Same Questionnaire: Using Large Language Models to Reliably Diversify Questions" Yun and colleagues (2024) explored the application of Large Language Models (LLMs) to generate varied versions of standardized questionnaires to mitigate fatigue.

Keeping Users Engaged During Repeated Interviews by a Virtual Agent: Using Large Language Models to Reliably Diversify Questions

Hye Sun Yun
yun.hy@northeastern.edu
Northeastern University
Boston, MA, USA

Mehdi Arjmand
arjmand.me@northeastern.edu
Northeastern University
Boston, MA, USA

Phillip Sherlock
phillip.sherlock@ufl.edu
University of Florida
Gainesville, FL, USA

Michael K. Paasche-Orlow
mpo@tufts.edu
Tufts University
Boston, MA, USA

James W. Griffith
jamesgriffith@uchicago.edu
University of Chicago
Chicago, IL, USA

Timothy Bickmore
t.bickmore@northeastern.edu
Northeastern University
Boston, MA, USA

<https://arxiv.org/pdf/2311.12707>

Results of their experiment with three groups: Original Scale, LLM Generated Scale Variants and LLM Variants+Commentary revealed:

- **Acceptable levels of convergent validity** of LLM items with original items ;
- Slightly lower but **reasonable internal consistency**;
- Both LLM groups had **higher compliance** with respondent interactions with data collection;
- LLM based commentary was viewed as **artificial and possibly distracting** from data collection.



The Promise of LLMs within the Pre-Data Collection Phase: AI as Collaborator/Assistant



Buskirk and colleagues (2025) examined how various zero-shot prompt components impacted the quality of survey items generated from ChatGPT 3.5.

Including the word "survey" in the prompt significantly increased the likelihood that generated items conformed to questionnaire design conventions and were written at more accessible reading levels.

However, the benefit of this term varied depending on how response formats were specified. Prompts that combined "survey" with "response options" yielded the strongest performance across quality indicators, though interaction effects showed that these benefits were not consistent across all conditions.

Prompts that used only the word "question" and omitted explicit guidance on response formats produced the lowest-quality outputs across nearly every outcome. These items were often open-ended, ambiguous, or structurally unsound.

The Task Is to Improve the Ask: An Experiment for Developing Prompts to Generate High Quality Survey Items from Large Language Models

22 Pages · Posted: 4 Aug 2025

[Trent Buskirk](#)
Old Dominion University

[Adam Eck](#)
Oberlin College

[Jerry Timbrook](#)
RTI International

Date Written: July 15, 2025

Abstract

Mentions of chatbots and generative artificial intelligence (AI) technologies—especially tools like ChatGPT—have become nearly ubiquitous in public discourse, prompting questions about their impact across a wide range of fields. Notably, Eloundou et al. (2023) identified survey research as one of the most affected industries, raising important questions about how these technologies might reshape workflows. In response, survey scientists

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5377878

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

Precautions for LLMs within the Pre-Data Collection Phase: AI as Collaborator/Assistant



Barends and de Vries (2024) use ChatGPT 4.0 to generate a short version of the HEXACO inventory and compare it to current abbreviated versions used in practice.

The results indicate that there were **no differences found between the ChatGPT short inventory compared to the abbreviated version used by humans in psychometric properties** such as internal consistency, convergent and discriminant validity and criterion-related validity.

However, GPT was **not able to discernably improve its original inventory on either internal consistency or content validity** even when instructed to. ChatGPT covered all facets per trait and was not able to modify this tendency, specifically.

ChatGPT made several mistakes when generating negatively keyed items.

Developing and Improving Personality Inventories Using Generative Artificial Intelligence: The Psychometric Properties of a Short HEXACO Scale Developed Using ChatGPT 4.0

Ard J. Barends  & Reinout E. de Vries

Pages 419-425 | Received 23 Sep 2024, Accepted 08 Dec 2024, Published online: 27 Dec 2024

“ Cite this article  <https://doi.org/10.1080/00223891.2024.2444454>



<https://www.tandfonline.com/doi/full/10.1080/00223891.2024.2444454>

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

Precautions for LLMs within the Pre-Data Collection Phase: AI as Collaborator/Assistant



Olson and Buskirk (2025) examine readability calculations for 60 survey questions performed by commonly available AI tools, at three time points over 2024-2025. We compare these to a “gold standard” online readability assessment tool (Readable.com), and calculations derived from R and Python functions.

The findings suggest that open-source tools are still the most robust methods for readability assessments at this point, even though LLM skills are evolving.

Some LLM models were prone to errors in identifying and correctly interpreting building blocks of readability equations, especially around monosyllabic words and word counting.

International Journal of Market Research

MRS Evidence Matters

Impact Factor: 1.4 / 5-Year Impact Factor: 3.3

Journal Home

Available access | Research article | First published online October 21, 2025 | Request permissions

“ChatBot” is a Two Syllable Word...Or Is It?: Using Generative AI for Survey Question Readability Assessments

Kristen Olson and Trent D. Buskirk

Volume 68, Issue 1 | <https://doi-org.proxy.lib.odu.edu/10.1177/14707853251389789> | View article versions

Contents | PDF/EPUB | Cite | Share options | Information, rights and permissions | Metrics and citation

Abstract

Market and survey researchers aim to write survey questions so that the target population can understand them. A common recommendation for general population studies is to write survey questions at an eighth-grade reading level. To evaluate whether questions meet this threshold, survey researchers turn to readability measures, such as the Flesch-Kincaid Reading Grade Level. Researchers may be able to streamline the calculation of question reading levels using artificial intelligence (AI) tools, such as ChatGPT, in

<https://journals.sagepub.com/doi/abs/10.1177/14707853251389789>



The Promise of LLMs within the Data Collection Phase: AI as Interviewer



Wuttke and colleagues (2024) recently investigated whether LLMs (ChatGPT 4) can effectively replace human interviewers to conduct scalable conversational interviews thereby balancing depth and scalability.

Utilized Humans and LLMs to conduct interviews based on predefined questionnaires covering political topics.

<https://arxiv.org/pdf/2410.01824>

AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers

Alexander Wuttke¹, Matthias Aßenmacher¹, Christopher Klamm²,
Max M. Lang³, Quirin Würschinger¹, Frauke Kreuter¹,

¹LMU Munich, ²University of Mannheim, ³University of Oxford,
Correspondence: a.wuttke@lmu.de

Abstract

Traditional methods for eliciting people's opinions face a trade-off between depth and scale: structured surveys enable large-scale data collection but limit respondents' ability to express

options have significant limitations (Schwarz and Hippler, 1987; Kash, 2013). Their static and impersonal nature often leads to respondent fatigue, which can diminish engagement and, consequently, the quality of responses (Krosnick, 1999; Jeong



Of all violations in interviewer behavior AI had a majority of the “ask follow-ups” and “don't be judgy” fails. Compared to text input, AI interviews may be longer but less elaborate.



Of all violations in interviewer behavior Humans had a majority of “active listening” fails.

The Promise of LLMs within the Data Collection Phase: AI as Interviewer



- Experimental evidence from Shoubik et al. (2025) further demonstrates that AI interviewers (via SmartProbe LLM) 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.
 - AI Chatbot was successfully able to identify when elaboration was needed.
 - The Chatbot assigned condition had slightly higher dropout after first probe compared to the equivalent point in the standardized condition, but those who remained reported only slightly less favorable experiences, even after receiving multiple probes by the end of the survey.

AI-Assisted Conversational Interviewing: Effects on Data Quality and Respondent Experience

Soubhik Barari, Jarret Angbazo, Natalie Wang, Leah M. Christian,
Elizabeth Dean, Zoe Slowinski, Brandon Sepulvado

Methodology & Quantitative Social Sciences
NORC at the University of Chicago

December 1, 2025

Abstract

Standardized surveys scale efficiently but sacrifice depth, while conversational interviews improve response quality at the cost of scalability and consistency. This study bridges the gap between these methods by introducing a framework for AI-assisted conversational interviewing. To evaluate this framework, we conducted a web survey experiment where 1,800 participants were randomly assigned to AI 'chatbots' which use large language models (LLMs) to dynamically probe respondents for elaboration and interactively code open-ended responses to fixed questions developed by human researchers. We assessed the AI chatbot's performance in terms of coding accuracy, response quality, and respondent experience. Our findings reveal that AI chatbots perform moderately well in live coding even without survey-specific fine-tuning, despite slightly inflated false positive errors due to respondent acquiescence bias. Open-ended responses were more detailed and informative, but this came at a slight cost to respondent experience. Our findings highlight the feasibility of using AI methods such as chatbots enhanced by LLMs to enhance open-ended data collection in web surveys.

<https://arxiv.org/pdf/2504.13908>

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

Precautions for LLMs within the Data Collection Phase: AI as Interviewer



Tirumala and colleagues (2025) examined the fitness for use of AI interviews compared to IVR technology in both quantitative and qualitative contexts.

AI interviews appear to provide more functionality than current IVR systems for quantitative survey projects but their overall utility for quantitative and qualitative projects may depend heavily on context and purpose:

- when human interviewers are unavailable;
- nuanced emotional detection is not essential
- topics are highly sensitive and prone to social desirability bias.

Published as a conference paper at COLM 2025

Mic Drop or Data Flop? Evaluating the Fitness for Purpose of AI Voice Interviewers for Data Collection within Quantitative & Qualitative Research Contexts

Shreyas Tirumala, Nishant Jain & Danny D. Leybzon
VKL Research, Inc.
San Francisco, CA
{shreyas,nishant,danny}@vkl.ai

Trent D. Buskirk
Professor, Provost Fellow of Data Science
Old Dominion University
Norfolk, Virginia
tbuskirk@odu.edu

Abstract

Transformer-based Large Language Models (LLMs) have paved the way for "AI interviewers" that can administer voice-based surveys with respondents in real-time. This position paper reviews emerging evidence to understand when such AI interviewing systems are fit for purpose for collecting data within quantitative and qualitative research contexts. We

Precautions for LLMs within the Data Collection Phase: AI as Interviewer



- 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,
 - 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 (Yu et al. 2024):
 - inconsistent probing and in varying order
 - privacy and data-security concerns [non-enterprise versions of Closed LLMs should be avoided]
 - potential biases stemming from opaque model training corpora

The Promise of LLMs within the Data Collection Phase: AI as Respondents



- Silicon samples/synthetic respondents represent a large opportunity but also a large threat to the validity of panels.
 - Here, large language models are used to generate sets of responses to survey instruments, often conditioned on demographic, attitudinal, or contextual information, to approximate how members of a target population might answer.

Argyle and colleagues (2023): Introduced the concept of “silicon samples” and criteria for assessing “algorithmic fidelity” for LLMs and report nuanced similarities between human and AI generated responses.

<https://bit.ly/ArgyleEtAl2023>



Political Analysis

Article contents



Abstract

Introduction

The GPT-3 Language

Out of One, Many: Using Language Models to Simulate Human Samples

Published online by Cambridge University Press: 21 February 2023

Lisa P. Argyle , Ethan C. Busby, Nancy Fulda, Joshua R. Gubler , Christopher Rytting and David Wingate

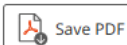
Show author details ▾

Article

Figures

Supplementary materials

Metrics



Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human subpopulations in social science research. Practical and research

**RANDOMNESS
BY DESIGN.**
AI TO REFINE.

The Promise of LLMs within the Data Collection Phase: AI as Respondents



Sun and colleagues (2023): Improved upon the concept of silicon samples by introducing so called “random silicon sampling” and showed it performed as well or better than silicon sampling for many tasks.

<https://arxiv.org/pdf/2402.18144>

Park and colleagues (2024) used the content of two-hour in-depth qualitative interviews of participants to create synthetic responses from LLMs. The respective synthetic respondents replicated the actual human respondent answers to a battery of GSS items 85% as accurately as humans replicated their own responses.

<https://arxiv.org/pdf/2411.10109>

Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information

Seungjong Sun¹, Eungu Lee¹, Dongyan Nan², Xiangying Zhao², Wonbyung Lee¹,
Bernard J. Jansen³, Jang Hyun Kim^{1,2}

¹Department of Applied Artificial Intelligence, ²Department of Interaction Science, Sungkyunkwan University

³Qatar Computing Research Institute, Hamad Bin Khalifa University

{tmdwhd406, dldmsrn0516, zxy94, co2797}@g.skku.edu, {ndyzxy0926, alohakim}@skku.edu, jjansen@acm.org

Abstract

Large language models exhibit societal biases associated with demographic information, including race, gender, and others. Endowing such language models

human-like biases associated with race, gender, ethnicity, and others from human-written data (Schramowski et al., 2023; Peters and Matz, 2023). Although many studies have attempted to mitigate societal biases in LLMs (Barocas and Selbst, 2016;

Generative Agent Simulations of 1,000 People

Authors: Joon Sung Park^{1*}, Carolyn Q. Zou^{1,2}, Aaron Shaw², Benjamin Mako Hill³, Carrie Cai⁴, Meredith Ringel Morris⁵, Robb Willer⁶, Percy Liang¹, Michael S. Bernstein¹

Affiliations:

¹Computer Science Department, Stanford University; Stanford, CA, 94305, USA.

²Department of Communication Studies, Northwestern University; Evanston, IL, 60208, USA.

³Department of Communication, University of Washington; Seattle, WA 98195, USA.

⁴Google DeepMind; Mountain View, CA 94043, USA.

⁵Google DeepMind; Seattle, WA 98195, USA.

⁶Department of Sociology, Stanford University; Stanford, CA, 94305, USA.

*Corresponding author. Email: joonspk@stanford.edu

Abstract:

The promise of human behavioral simulation—general-purpose computational agents that replicate human behavior across domains—could enable broad applications in policymaking and social science. We present a novel agent architecture that simulates the attitudes and behaviors of 1,052 real individuals—applying large language models to qualitative interviews about their

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

The Promise of LLMs within the Data Collection Phase: AI as Respondents



Ehrett and colleagues (2024) explore the use of open LLMs (Llama and Alpaca) for data augmentation for a specific text-classification task.

Using open-ended responses collected as part of a small survey to hospital staff, LLMs were used to generate synthetic data like those collected within the survey.

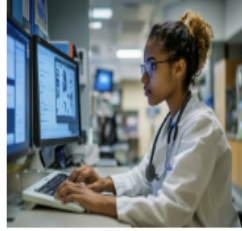
A screenshot of a webpage from JMIR Medical Education. The page features a blue header with the journal name and navigation options. Below the header, there is a publication date and a link to preprints. The main content area includes a photograph of a healthcare professional at a computer, the title of the article, and the names of the authors with their ORCID iD icons.

JMIR Medical Education






Journal Information ▾ Browse Journal ▾

Published on 19.Nov.2024 in Vol 10 (2024)

Preprints (earlier versions) of this paper are available at <https://preprints.jmir.org/preprint/51433>, first published 31.Jul.2023.



Leveraging Open-Source Large Language Models for Data Augmentation in Hospital Staff Surveys: Mixed Methods Study

Carl Ehrett¹ ; Sudeep Hegde² ; Kwame Andre³ ; Dixizi Liu² ; Timothy Wilson² 

Leveraging the survey data and labelled synthetic data, the team then applied another set of LLMs for classifying responses into thematic categories.

Evaluation included both the data creation and classification tasks taken together as a two-step approach. Llama 7B followed by RoBERTa was the optimal combination of all tested (AUC=0.87).



Precautions for LLMs within the Data Collection Phase: AI as Respondent



- Unlike classical imputation or synthetic data methods based on explicit probabilistic models (e.g., multiple imputation by chained equations or predictive mean matching), these responses emerge from (Argyle et al. 2025) :
 - an opaque combination of prompts
 - patterns learned from massive but not necessarily representative training corpora
 - post-training alignment procedures
- The validity of these samples, or inferences derived using them, depend on an aggregate effect of these opaque components – making error or quality very difficult to quantify.
- von der Heyde (forthcoming, 2026) is tracing *new* errors that are introduced by incorporating LLMs into the various stages of the survey research process [Coming to AAPOR, 2026].



Precautions for LLMs within the Data Collection Phase: AI as Respondent



Bisbee and colleagues (2023) report contrary findings that suggest that silicon samples 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 version being used.

<https://bit.ly/BisbeeEtAl2024>

PA
POLITICAL ANALYSIS

Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

Published online by Cambridge University Press: 17 May 2024

James Bisbee , Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson Show author details

Article Figures Supplementary materials Metrics

Save PDF Share Cite Rights & Permissions

Abstract

Large language models (LLMs) offer new research possibilities for social scientists, but their potential as “synthetic data” is still largely unknown. In this paper, we investigate how accurately the popular LLM ChatGPT can recover public opinion, prompting the LLM to adopt

Santurkar and colleagues (2023) provide evidence of a **lack of representation of opinions from older widowed women** in LLM output among others. They also found that language models tuned with reinforcement learning from human feedback are **more aligned with left-leaning, liberal views**.

Whose Opinions Do Language Models Reflect?

Shibani Santurkar¹ Esin Durmus¹ Faisal Ladhak² Cino Lee¹ Percy Liang¹ Tatsunori Hashimoto¹

Abstract

Language models (LMs) are increasingly being used in open-ended contexts, where the opinions they reflect in response to subjective queries can improve the model, to the model designers themselves. This motivates the central question of our work:

Whose opinions (if any) do language models reflect?

<https://proceedings.mlr.press/v202/santurkar23a/santurkar23a.pdf>

AI TO REFINE.

Precautions for LLMs within the Data Collection Phase: AI as Respondent



- Westwood (2025) demonstrates that autonomous AI agents, operating from a simple prompt, can evade current detection methods and produce high-quality survey responses that demonstrate reasoning and coherence expected of human responses.
- This capability illustrates possible compromises to the integrity of responses from panelists or survey participants, writ large under current monitoring conditions.
- These responses while synthetic, represent a deeper threat in that they are unintentionally gathered and offered by potentially bad actors – and can adversely steer results.

PNAS RESEARCH ARTICLE | POLITICAL SCIENCES OPEN ACCESS Check for updates

The potential existential threat of large language models to online survey research

Sean J. Westwood^{a1}

Edited by James N. Druckman, University of Rochester, Rochester, NY; received July 9, 2025; accepted September 12, 2025 by Editorial Board Member Margaret Levi

The advancement of large language models poses a severe, potentially existential threat to online survey research, a fundamental tool for data collection across the sciences. This work demonstrates that the foundational assumption of survey research—that a coherent response is a human response—is no longer tenable. I designed and tested an autonomous synthetic respondent capable of producing survey data that possesses the coherence and plausibility of human responses. This agent successfully evades a comprehensive suite of data quality checks, including instruction-following tasks, logic puzzles, and “reverse shibboleth” questions designed to detect nonhuman actors, achieving a 99.8% pass rate on 6,000 trials of standard attention checks. The synthetic respondent generates internally consistent responses by maintaining

Significance

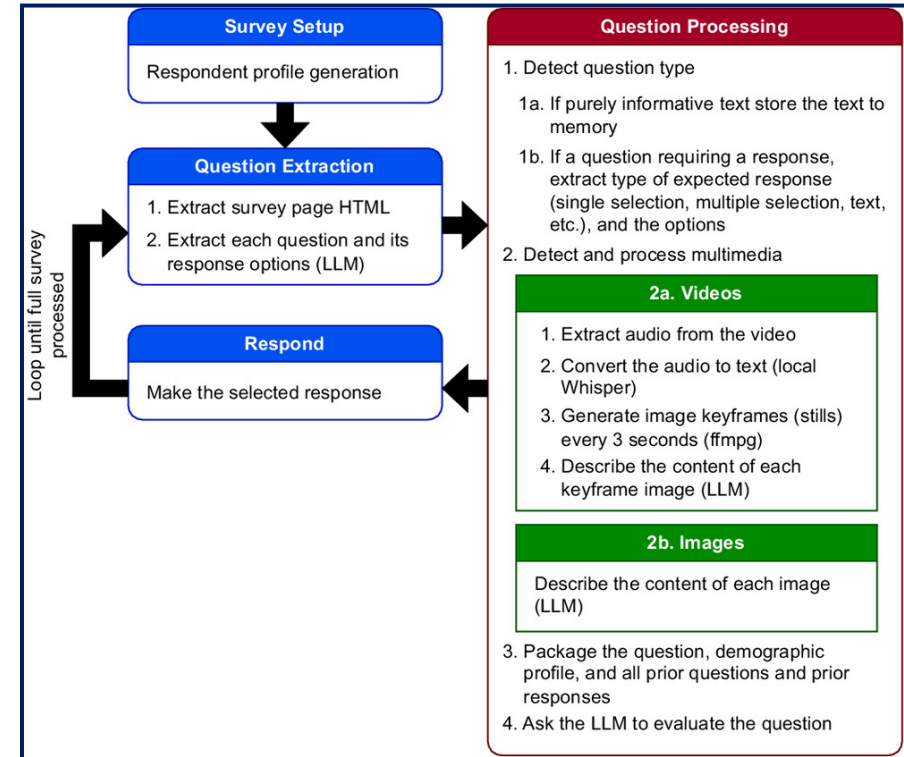
Surveys are a primary source of data across the sciences, from medicine to economics. I demonstrate that the assumption that logically coherent responses are from humans is now untenable. I show that

Attribution: <https://doi.org/10.1073/pnas.2518075122>

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

Sizing the Threat of Autonomous AI Responses

- To appreciate the extent to which the threats demonstrated by Westwood (2025) requires an understanding of the architecture of the generation process.
- Using two-layer design autonomous agents first extract the questions and then respond to them.
- While such a system could be run on a personal computer the question about whether “typical” panel participants would actually use such tools (or even develop them) for their use on the panel is outstanding.



Attribution: Figure 1, Westwood (2025):
<https://doi.org/10.1073/pnas.2518075122>

Precautions for LLMs within the Data Collection Phase: AI as Respondent



Companies are now offering and specializing in providing synthetic samples (or a mix of them with human responses) to researchers at costs that are far cheaper than fielding/obtaining probability samples either outright or from probability panel providers.

Expected Parrot

Products ▾ Company Resources ▾ Log in Try now

Simulate your customers

Expected Parrot is an open-source framework and an interactive app to design, test, and simulate agents for surveys and research.

Work in code → Work interactively →

ACCELERATING RESEARCH AT LEADING INSTITUTIONS

MIT Massachusetts Institute of Technology HARVARD UNIVERSITY Stanford Berkeley UNIVERSITY OF OXFORD COLUMBIA UNIVERSITY Yale

<https://www.expectedparrot.com/>

qualtrics^{XM}
DIGITAL SUCCESS

About Synthetic Panels

Recruiting the right participants for a study can be difficult. You may not get the exact demographics you need, and the shorter the deadline, the less sure you can be that everyone will answer on time. One possible solution can be to use synthetic panels.

Synthetic panels are powered by a first party proprietary AI model developed here at Qualtrics. Our synthetic panel is trained on thousands of responses from a variety of demographic backgrounds in order to more accurately predict how certain populations would respond to a survey.

Our synthetic panel is based on the **United States General Population**, and is only available in **English**. This panel comes with ready-made quotas and target breakouts in order to represent your chosen population and make it easy to launch your survey right away.

<https://www.qualtrics.com/support/survey-platform/distributions-module/synthetic-panels/>

The Promise of LLMs within the Post-Data Collection Phase: AI as Pre-Processor



Allamong and colleagues (2025) used ChatGPT 4o with different zero-shot prompts to correct misspellings in open-ended survey responses from ANES data from 1996 to 2020.

GPT was shown to correct 85-90% of the misspellings identified in a random subsample by human coders and outperformed R software designed to do the same task.

The authors show that failure to correct misspellings at scale could result in significant underestimates in the amount of emotionality present in the responses.

Research & Politics
Volume 12, Issue 1, January 2025
© The Author(s) 2025, Article Reuse Guidelines
<https://doi-org.proxy.lib.edu/10.1177/20531680241311510>

Sage Journals

Research Article



Spelling correction with large language models to reduce measurement error in open-ended survey responses

Maxwell B. Allamong¹, Jongwoo Jeong², and Paul M. Kellstedt ³

Abstract

Open-ended survey questions have a long history in public opinion research and are seeing a renewed interest as computing power and tools of text analysis proliferate. A major challenge in performing text analyses on open-ended responses is that the documents—especially if transcribed or collected through web surveys—may contain measurement error in the form of misspellings which are not easily corrected in a reliable and systematic manner. This paper provides evidence that large language models

<https://journals.sagepub.com/doi/10.1177/20531680241311510>

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

The Promise of LLMs within the Post-Data Collection Phase: AI as Labeler/Classifier



Törnberg (2023) compared MTurk Qualified workers with Chat-GPT 4 (at two temperatures) and two political science experts for coding political party of 500 political candidates based on posted tweets. Accuracy of ChatGPT was above 90% compared to between 80 and 85%, on average for the human coders.

Gilardi and colleagues (2023) compared the accuracy of MTurkers and ChatGPT 3.5 on different annotation tasks (relevance, stance, topics and frame detection) using a sample of Tweets and News Articles.

- Authors found that ChatGPT zero-shot accuracy exceeded that of MTurkers for almost all tasks and higher intercoder agreement for all tasks.
- Cost per annotation for ChatGPT was \$0.003, 30 times cheaper than MTurk.

<https://www.pnas.org/doi/epdf/10.1073/pnas.2305016120>

ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning

Petter Törnberg^{a,c,1}

^aAmsterdam Institute for Social Science Research (AISSR), University of Amsterdam

This manuscript was compiled on April 14, 2023

This paper assesses the accuracy, reliability and bias of the Large Language Model (LLM) ChatGPT-4 on the text analysis task of classifying the political affiliation of a Twitter poster based on the content of a tweet. The LLM is compared to manual annotation by both expert classifiers and crowd workers, generally considered the gold standard for such tasks. We use Twitter messages from United States politicians during the 2020 election, providing a ground truth against which to measure accuracy. The paper finds that ChatGPT-4 has achieved higher accuracy, higher reliability, and equal or lower bias than the human classifiers. The LLM is able to correctly annotate messages that require reasoning on the basis of contextual know-

Language Processing and Machine Learning – have evolved quickly in recent years (6), common methods that build on bag-of-words, lexical meanings and sentence semantics tend to work poorly for tasks requiring complex inferences on the basis of knowledge about the world or assumptions of the author's mental state and intentions.

The recent rise of generative AI may however be at the cusp of changing this. In recent months, ChatGPT has become a global sensation, and one of the quickest growing consumer products of all time. ChatGPT is a pre-trained model based on a massive neural network with billions of parameters, and trained on knowledge of billions of words from the T-

13 Apr 2023

<https://arxiv.org/abs/2304.06588>

PNAS

BRIEF REPORT | POLITICAL SCIENCES

OPEN ACCESS



ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi^{a,1}, Meysam Alizadeh^{a,1}, and Maël Kubli^{a,1}

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using four samples of tweets and news articles ($n = 6,183$), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than \$0.003—about thirty times cheaper than MTurk. These results demonstrate

Precautions for LLMs within the Post-Data Collection Phase: AI as Labeler/Classifier



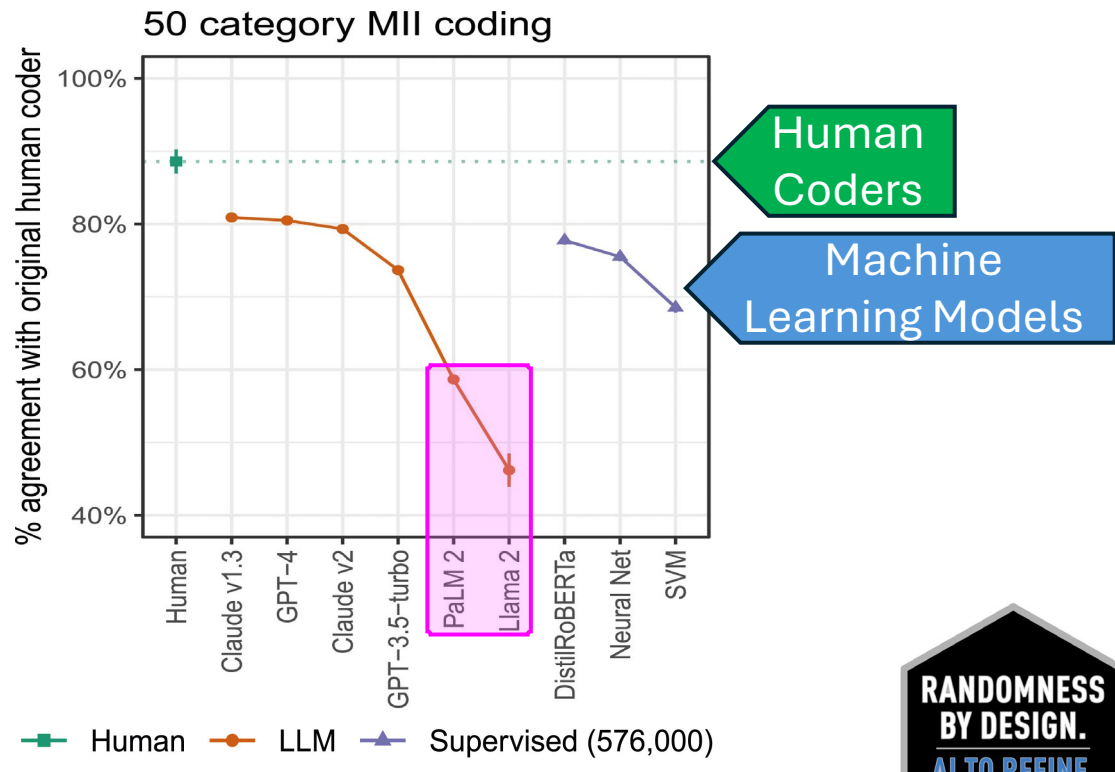
Mellon and colleagues (2024) compared the accuracy of six LLMs (with few-shot prompting) to supervised ML and human coders for categorizing the “most important issue” using data from the British Election Study Internet Panel.

- The LLMs did not all perform sufficiently well!
- For larger LLM models (i.e. Claude-1.3) the performance was just behind human coders.
 - While larger LLMs were functionally close to human performance, smaller, open-sourced LLMs fell behind by nearly 20 percentage points or more.
 - In particular, PaLM-2 and Llama-2 were worse across all metrics and scenarios.

Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale

[Jonathan Mellon](#) [Jack Bailey](#) [\[...\]](#) and [Phillip Schmedeman](#) [View all authors and affiliations](#)

[All Articles](#) | <https://doi-org.proxy.lib.ou.edu/10.1177/20531680241231468>



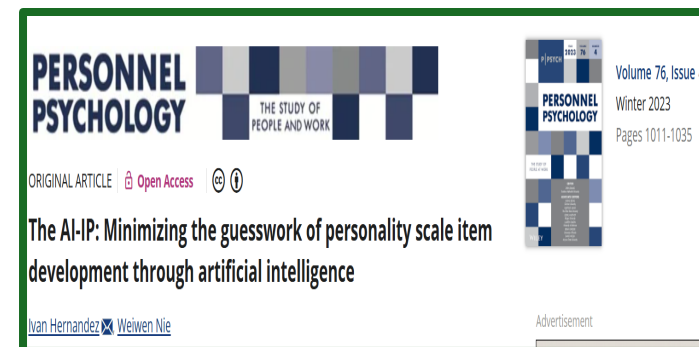


Leveraging Panel Resources with LLMs

- Probability panels are the gold standard for data collection and being able to further leverage that strength with great questions is paramount to maximizing total quality.
- One resource panels have is lots of prior survey items. What if you could leverage this rich resource as a way to jump start generation of new items?

Leveraging LLMs to be trained on such a bank of items for the express purpose of generating many more items is possible and a framework for this approach has been reported by Hernandez and Nie (2022).

- They applied a Chat-GPT 2 model to learn the structure of personality items from 3300 plus items from the International Personality Item Pool to generate over 1,000,000 new items (that were about 1 word shorter, on average and were about 6 points lower on Flesch Reading Ease metrics).
- They also used embeddings from another model to determine which items were proximal to other items with high degrees of accuracy.



Leveraging Panel Resources with LLMs...

- Think about the value of public-release tables and possibly custom tables that could be used for TABULAR LLM development and testing – See Wu, Ritter and Xu (2025), for examples.
 - Not many such “benchmark” data exist for public opinion or behavioral data.
- Fine tuning of LLMs using current information stored in tables from probability panels could be leveraged if these tables were first serialized – made more AI readable with natural language representation (see Hegselmann et al., 2023).

**Tabular Data Understanding with LLMs:
A Survey of Recent Advances and Challenges**

Xiaofeng Wu Alan Ritter Wei Xu
College of Computing, Georgia Institute of Technology
xwu414@gatech.edu, alan.ritter@cc.gatech.edu, wei.xu@cc.gatech.edu

Abstract

Tables have gained significant attention in large language models (LLMs) and multimodal large language models (MLLMs) due to their complex and flexible structure. Unlike linear text inputs, tables are two-dimensional, encompassing formats that range from well-structured database tables to complex, multi-layered spreadsheets, each with different purposes. This diversity in format and purpose

31 Jul 2025

<https://arxiv.org/pdf/2508.00217>

17 Mar 2023

TabLLM: Few-shot Classification of Tabular Data with Large Language Models

Stefan Hegselmann^{1,2} Alejandro Buendia¹ Hunter Lang¹ Monica Agrawal¹ Xiaoyi Jiang² David Sontag¹
¹ MIT CSAIL ² University of Münster

Abstract

We study the application of large language models to zero-shot and few-shot classification of *tabular data*. We prompt the large language model with a serialization of the tabular data to a natural-language string, together with a short description of the classification problem. In the settings with a small number of training examples, i.e. the *few-shot* setting. While deep learning has led to breakthroughs in computer vision and natural language processing, this success has not yet been extended to the tabular domain. For example, self-supervised deep learning methods have been introduced for tabular data (Yin et al., 2020; Arik and Pfister, 2021), but Grinsztajn et al. (2022) showed that these deep techniques

<https://arxiv.org/pdf/2210.10723>

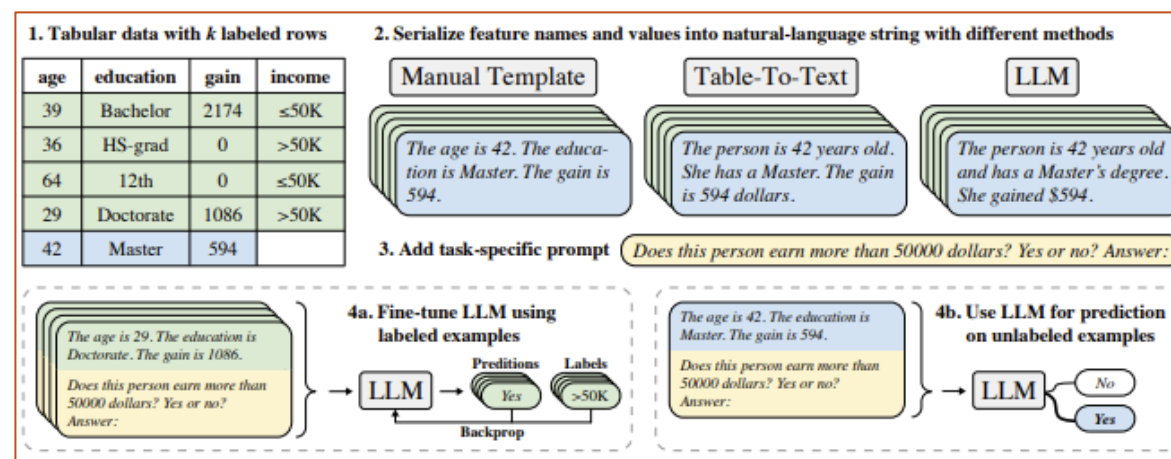
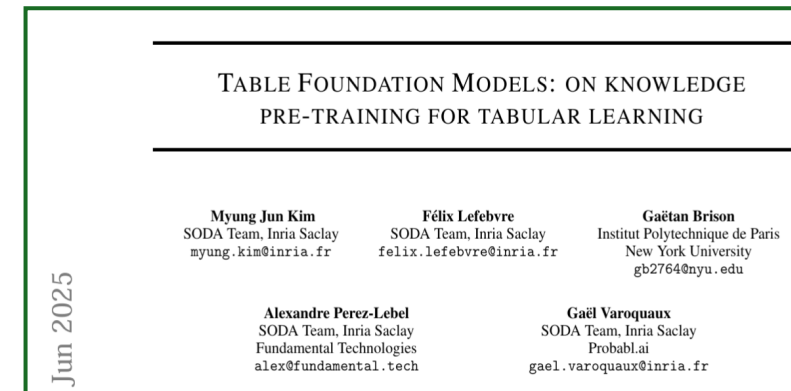


Figure 1 from Hegselmann et al., 2023



Panel's Learning from Panel Data

- Molnar (2026) and Kim et al. (2025) discuss the current state of **Tabular Foundation Models** which are AI models that are trained on Tables as input.
 - These models learn structure within tables and then apply that learning to infer structure in new tables.
- If Panels prepared tables of imputations used over time as part of the training data, these TFMs could then apply that learning to this year's data table to complete the imputation task.
- With an explainable layer added on to these models (explainable AI) you could learn more about what variables have driven trends over time, for example.
- A TFM trained on regular tables from the panel's paradata could also be another way to learn more about response, participation and attrition.



<https://arxiv.org/pdf/2505.14415>



Rapid Focus Grouping at Scale

<https://arxiv.org/pdf/2309.03220>

Conversational Swarm Intelligence, a Pilot Study

Louis Rosenberg
Unanimous AI
Pismo Beach, CA
louis@unanimous.ai

Gregg Willcox
Unanimous AI
Seattle, Washington
gregg@unanimous.ai

Hans Schumann
Unanimous AI
San Francisco, CA
hans@unanimous.ai

Miles Bader
Vassar College
Poughkeepsie, NY
mbader@vassar.edu

Ganesh Mani
Carnegie Mellon University
Pittsburgh, PA
ganeshm@andrew.cmu.edu

**Kokoro Sagae, Devang Acharya,
Yuxin Zheng, Andrew Kim,
and Jialing Deng**
Carnegie Mellon University

- Imagine being able to leverage panels for representative consensus round emerging topics for which no common questions yet exist because the topic is new and emergent.
- Rosenberg and colleagues (2025) use Conversational Swarm Intelligence from Biology along with LLMs to conduct rapid group-based insights from open-ended questions whose answers are not known in advance.
- The output of such systems not only capture the preferences of the group but also through conversational analysis, the reasons supporting or refuting those preferences.
- In pilot testing the results indicate that compared to traditional chat rooms this method can increase contributions from the group as a whole and reduce the gap between the most and least verbose group members.
 - Results also indicate that for some studies involving political candidate preferences, group insight convergence, with groups as large as 81 participants, can be achieved in a matter of minutes.

**RANDOMNESS
BY DESIGN.
AI TO REFINE.**

A Prob-**AI**-lity Panel for our Collective Futures...

- What if a probability panel could be constructed for multiple AI and societal purposes including:
 - Key set of national benchmarks – that could evolve over time and be used as possible sources of alignment for LLM models to minimize biases
 - Track people’s use of AI and possible job displacements at national/regional level
 - Possible source of representative reinforcement learning for LLM models (currently we have seen that reinforcement with convenience samples portends liberal leaning biases)
 - A means for data to seed larger predictive models we could use for imputation and other predictive tasks
 - this could be standardized, more transparent and perhaps adopted as one methodological approach multiple sources use
 - A means for rapid, in-depth interviews to be conducted and analyzed quickly and at scale
 - As a source for augmenting training data for LLMs into the future –
 - Making datasets more readable to machines and with topline reports that can be discovered in training sets and retrievals.



Planning for a Better Future, Together!

- **Experiment with LLMs within your workflow**
 - Sharing fitness for use of LLMs that can improve panel efficiency could be used by all as a way to ensure a more stable future for our gold standard of probability sampling.
- **Consider submitting your work to upcoming special issue of POQ dedicated to exploring AI in survey research and public opinion.**
 - **Co edited by Link, Kreuter and Buskirk**
 - **More info: <https://lnkd.in/eKCUxvWd>**



Planning for a Better Future, Together!

- Probability sampling is generally transparent with known/computable selection probabilities and designs publicly reported.
 - We don't want to weaken the transparency of our gold standard method with LLM-based approaches to sustain, manage or infer from such samples that can't be described reasonably.
- Consider using the forthcoming tool for reporting how and what you did in terms of LLM use in a standardized way to help pave the way for our future!

<https://hannahcha417.github.io/aapor-disclosure-checklist>

AI Disclosure Checklist




Guest Mode

For each question in this checklist, the researcher should indicate one of the following: ▲

- The answer to the question
- Question is not applicable
- Answer is unknown to the researcher and explain why
- Answer is proprietary and explain why
- Answer would violate privacy of participants and explain why




Immediate Disclosures

Immediate disclosures must be included in any reporting or methodological summaries and presented in a way that is clearly disclosed and easily accessible to readers.

1. Tasks Performed by AI 
2. Human Oversight or Validation 
3. Human Respondents Disclosure 

Core/Enhanced Questions

Core questions should be answered in all reporting scenarios ensuring consistent transparency across studies: this is the minimum viable disclosure to ensure that consumers of the polling data can understand potential bias and limitations. Enhanced questions are always valuable to answer, as they provide deeper insight into methods and AI involvement, but they are not mandatory in every situation: this is necessary for any situation that requires reproducibility.

4. Model Details 
5. Access/Tooling Details 
6. Core Prompts or Instructions 
7. Additional Enhanced Disclosures

Format: File Type:

THANK YOU!

Questions?

Comments?

tbuskirk@odu.edu

Please see me afterwards
for a commemorative
Conference Pin highlighting
CIPHER and this Talk!



References

- Rothschild, David M. and Brand, James and Schroeder, Hope and Wang, Jenny, Opportunities and risks of LLMs in survey research (October 28, 2024). Available at SSRN: <https://ssrn.com/abstract=5001645> or <http://dx.doi.org/10.2139/ssrn.5001645>
- Rothschild, D. M., Buskirk, T. D., Eckman, S., Hillygus, D. S., Kreuter, F., & Lazer, D. (2025). Successfully Navigating the Disruption AI will Bring to Survey Research. Available at: https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2025_July_N92_04.pdf
- Barari, S., Lerner, J.Y., Yan, T. and Christian, L.M. (2025) Generative AI in Survey Research: Principles and Use Cases. Paper presented at the 80th Annual American Association for Public Opinion Research Conference; St. Louis, MO, May 15, 2025. Available at: [Barari -- AAPOR presentation.pdf](#)
- Buskirk, T. D., Keusch, F., von der Heyde, L., & Eck, A. (2025a). More Parameters Than Populations: A Systematic Literature Review of Large Language Models within Survey Research. arXiv preprint arXiv:2509.03391. <https://doi.org/10.48550/arXiv.2509.03391>
- Buskirk, T., Eck, A., & Timbrook, J. (2025). The Task Is to Improve the Ask: An Experiment for Developing Prompts to Generate High Quality Survey Items from Large Language Models. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5377878.
- Kim, M. J., Lefebvre, F., Brison, G., Perez-Lebel, A., & Varoquaux, G. (2025). Table foundation models: on knowledge pre-training for tabular learning. *arXiv preprint arXiv:2505.14415*.
- Rosenberg, L., Willcox, G., Schumann, H., Bader, M., Mani, G., Sagae, K., ... & Deng, J. (2023). Conversational Swarm Intelligence, a Pilot Study. arXiv preprint arXiv:2309.03220.

COMMENT | 09 February 2026

How to deal with the survey-taking AI agents that threaten to upend social science

Researchers need new bot-detection strategies that exploit the limits of human reasoning rather than AI weaknesses.

By [Folco Panizza](#), [Yara Kyrchenko](#) & [Jon Roozenbeek](#)

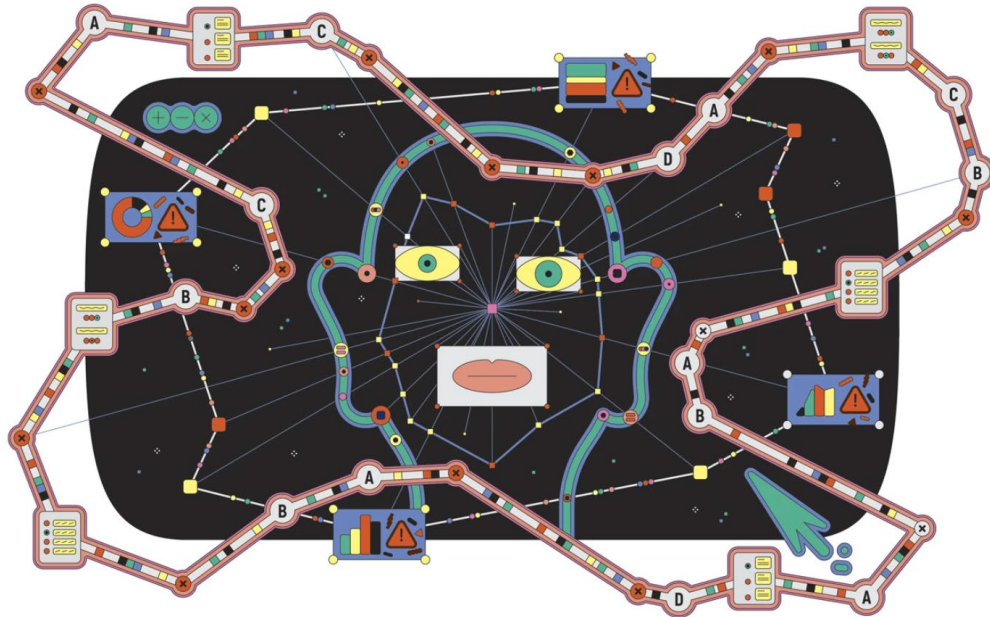


Illustration: Bratislav Milenković

<https://www.nature.com/articles/d41586-026-00386-2>

AI Disclosure Checklist

Guest Mode

For each question in this checklist, the researcher should indicate one of the following:

- The answer to the question
- Question is not applicable
- Answer is unknown to the researcher and explain why
- Answer is proprietary and explain why
- Answer would violate privacy of participants and explain why

Immediate Disclosures

Immediate disclosures must be included in any reporting or methodological summaries and presented in a way that is clearly disclosed and easily accessible to readers.

1. Tasks Performed by AI

2. Human Oversight or Validation

3. Human Respondents Disclosure

Core/Enhanced Questions

Core questions should be answered in all reporting scenarios ensuring consistent transparency across studies: this is the minimum viable disclosure to ensure that consumers of the polling data can understand potential bias and limitations. Enhanced questions are always valuable to answer, as they provide deeper insight into methods and AI involvement, but they are not mandatory in every situation: this is necessary for any situation that requires reproducibility.

4. Model Details

5. Access/Tooling Details

6. Core Prompts or Instructions

7. Additional Enhanced Disclosures

Format:

List

File Type:

PDF