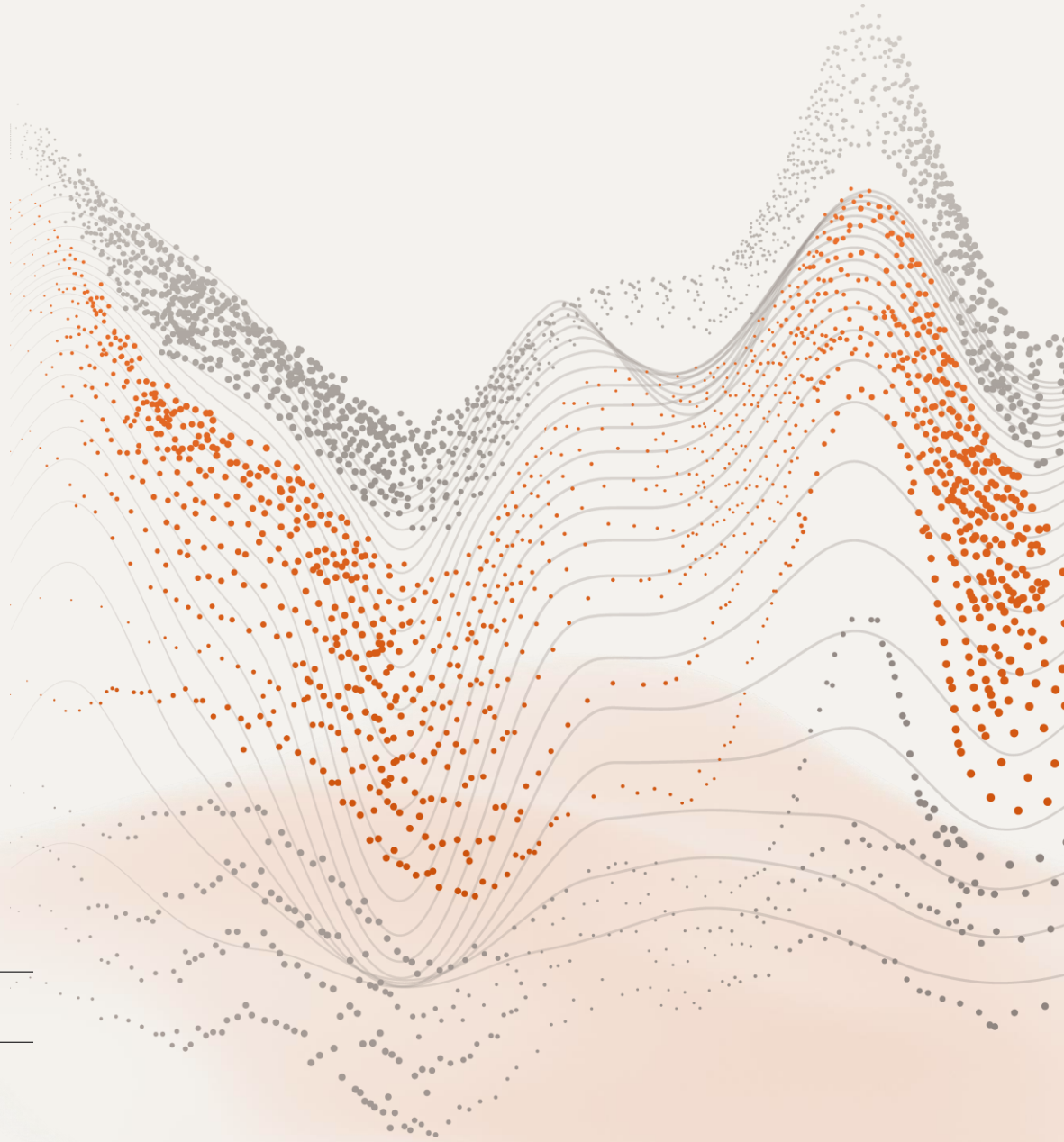
NORC Research Science

Responsible AI to Advance Survey Design and Data Collection: A Framework-Driven Approach

FedCASIC 2026

04/21/2026

Ting Yan



RESEARCH ARTICLE | POLITICAL SCIENCES |

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

Sean J. Westwood Authors Info & Affiliations

AI Doesn’t Threaten Surveys

Irresponsible Use of AI Does

The real risk is letting AI decide for us

- **AI is a general-purpose technology**
- **Human-in-the-loop is necessary but not sufficient**
- **Survey frameworks must govern AI use**

Existing frameworks are useful descriptors, not decision rules

Barari et al. (2025)

- AI as “Assistant”
- AI as “Methodology”
- AI as “Tool”

Rothschild et al. (2025)

- AI as the Research Colleague
- AI as the Interviewer
- AI as the Respondent
- AI as a Labeler
- AI as the Modeler
- AI as the Briefer

Buskirk et al. (2025a)

- Pre-data Collection Phase
- Data Collection Phase
- Post-Data Collection Phase

As a survey methodologist, I propose and recommend informing and governing AI use within survey frameworks

- **Survey Life Cycle**
- **Total Survey Error (TSE)**

Informing and governing the use of AI within survey frameworks

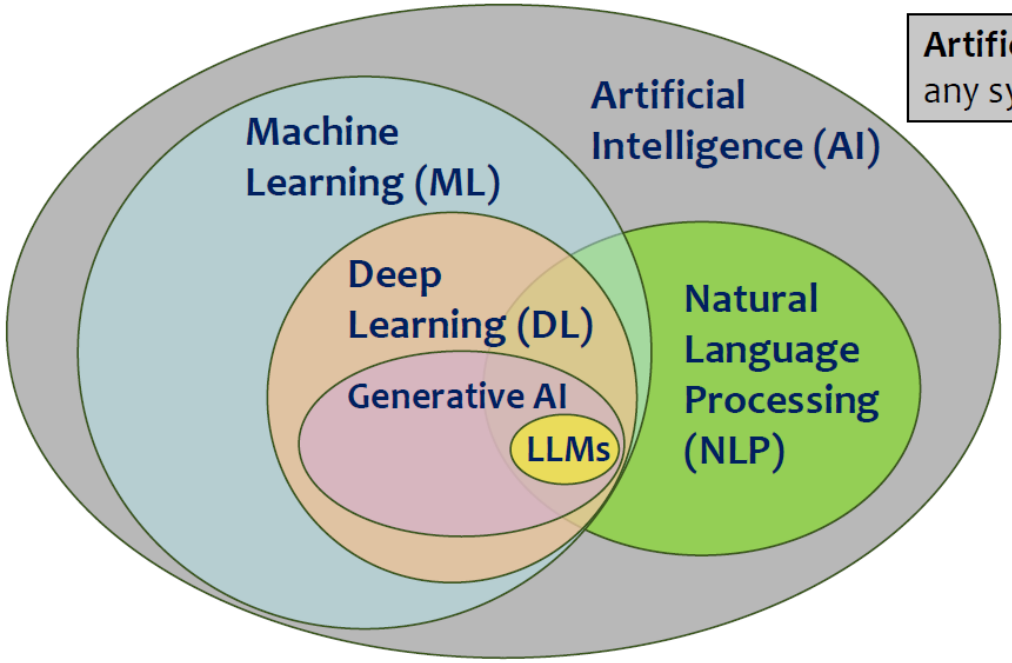
AI ROLES → SURVEY NEEDS

WHAT AI CAN DO → WHAT SURVEYS REQUIRE

Buskirk (2026)

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.

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.

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.

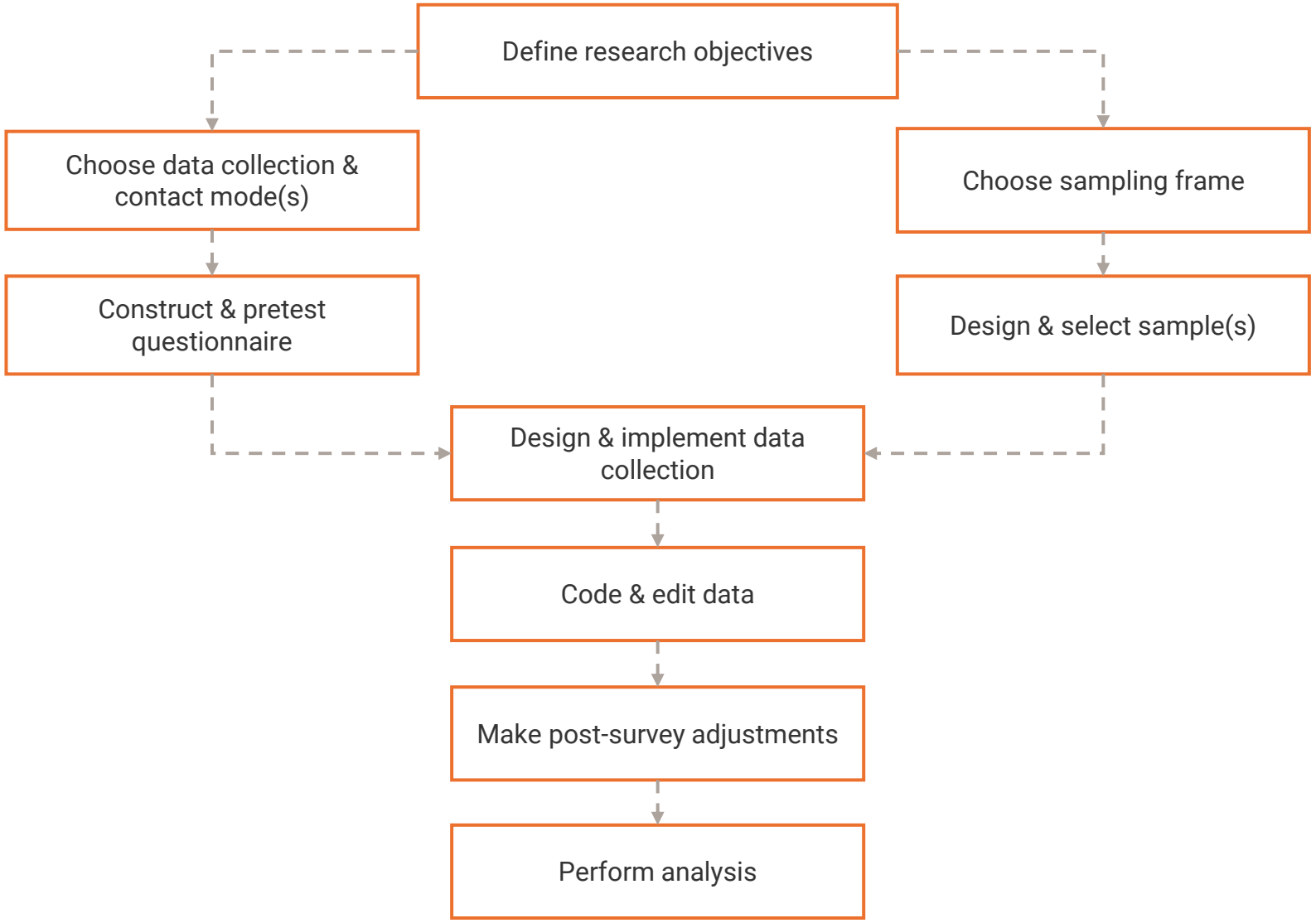
Responsible AI

- **Intentional**
- **Fit-for-purpose**
- **Human judgment and accountability**
- **TSE-guided validation and evaluation**
- **Transparent**

Governing AI with Survey Frameworks

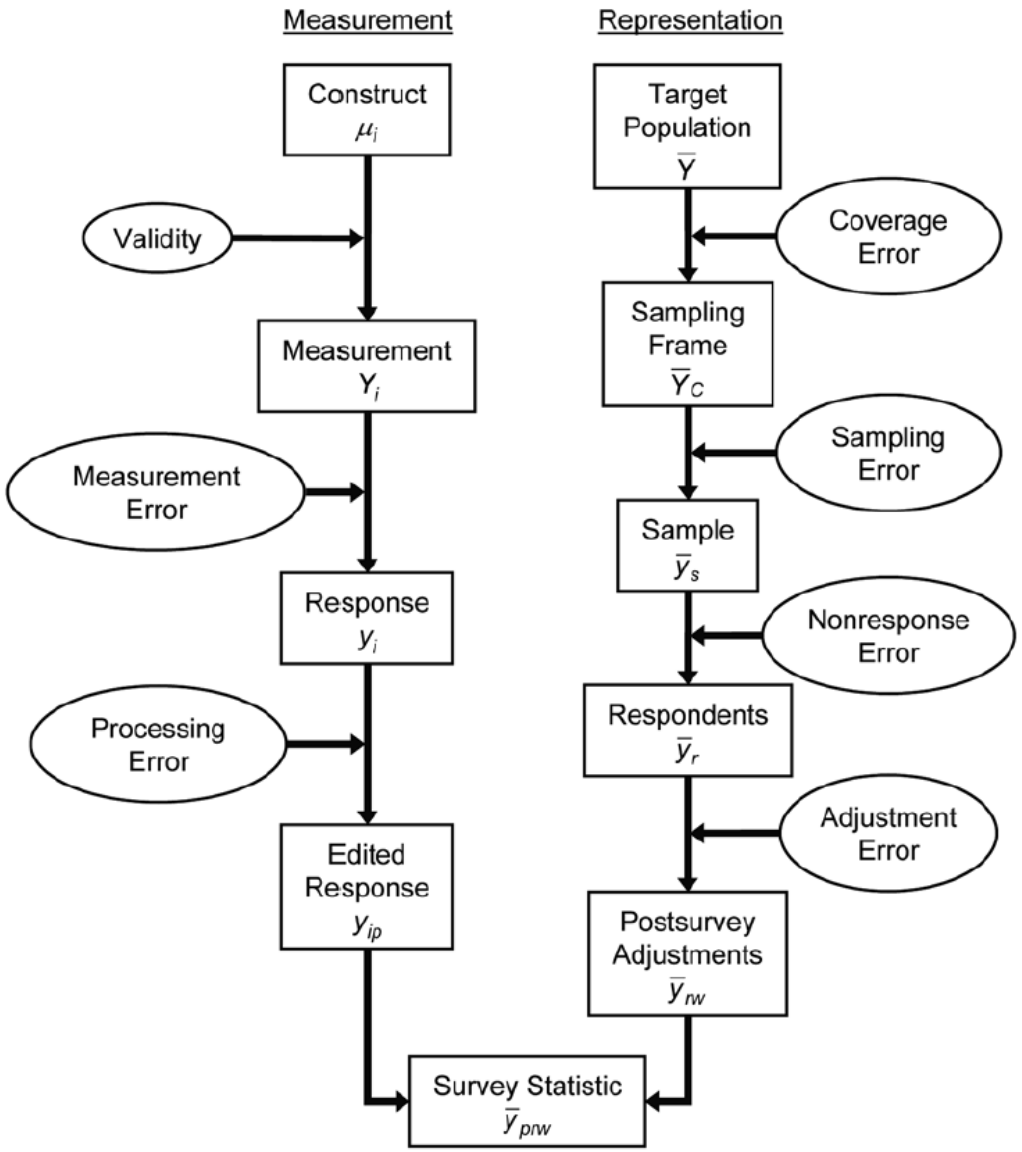
NORC Research Science

Use the survey life cycle to identify bottlenecks where AI may improve efficiency and/or quality



Adapted from Groves et al. 2009. *Survey methodology*. Hoboken, N.J.: Wiley.(Figure 5.1)

The TSE framework helps assess both error reduction and new error sources introduced by AI



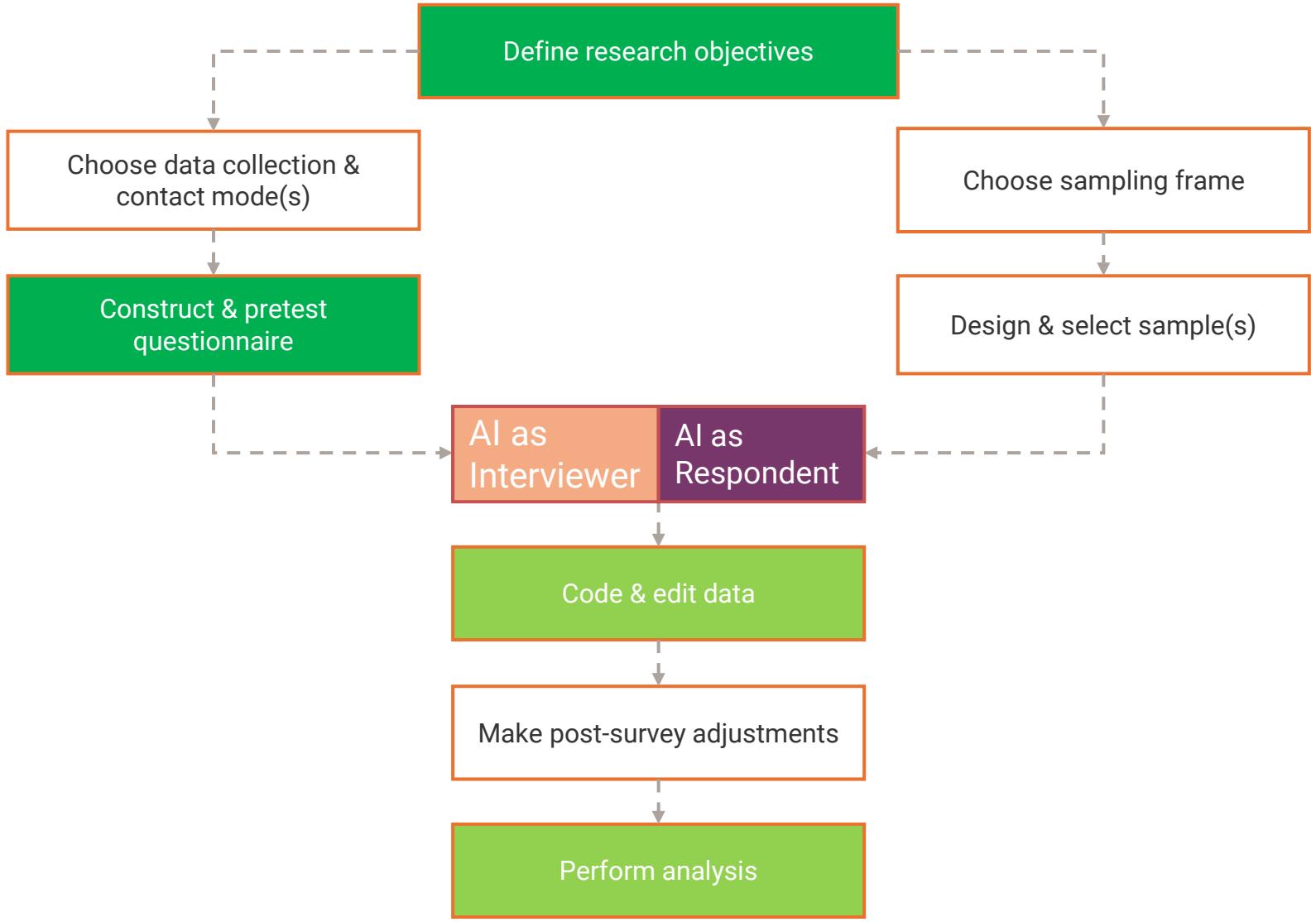
Groves et al. 2009. *Survey methodology*. Hoboken, N.J.: Wiley.

High potential, Manageable Risk

High potential, Medium Risk

High potential, High Risk

Limited Impact, High Risk



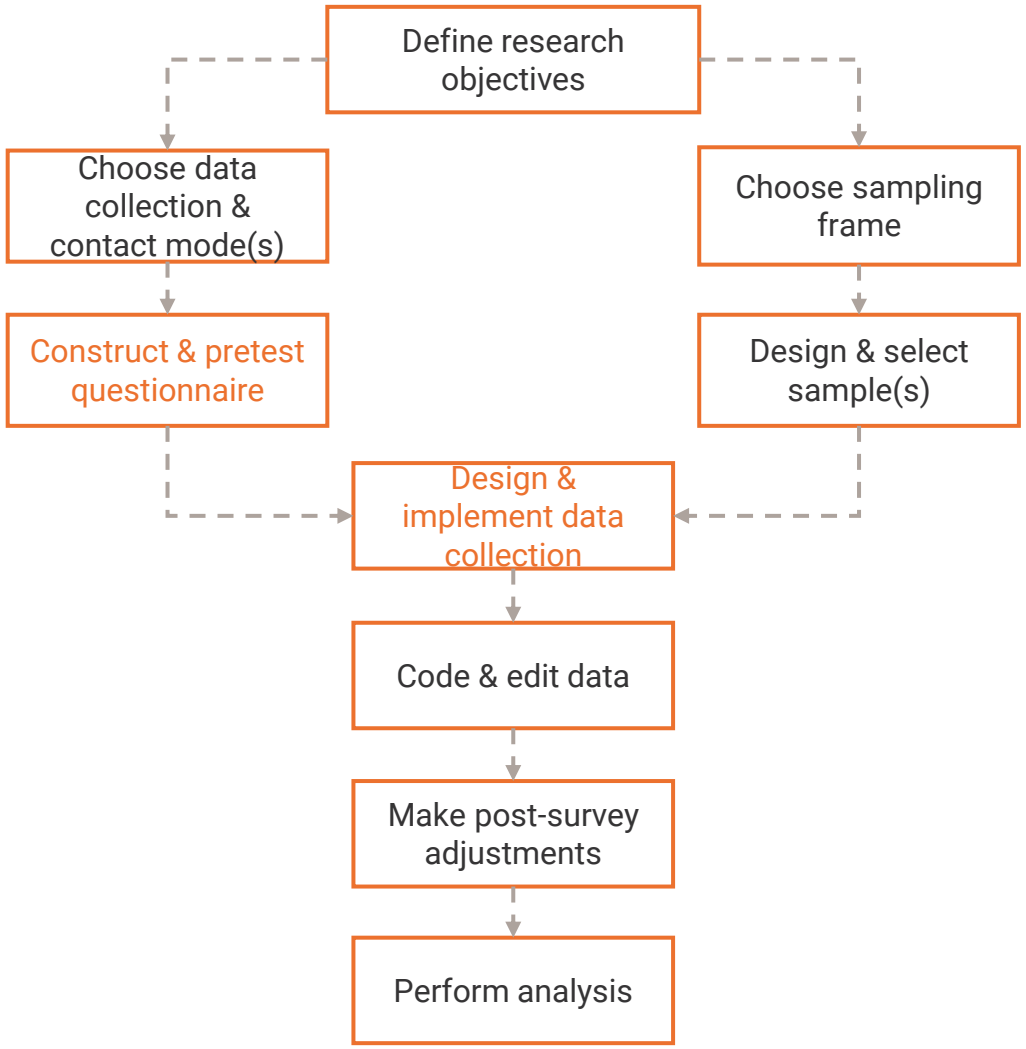
Rothschild et al. (2025)

Example 1: Use of AI to Improve Efficiency of Recordings Data

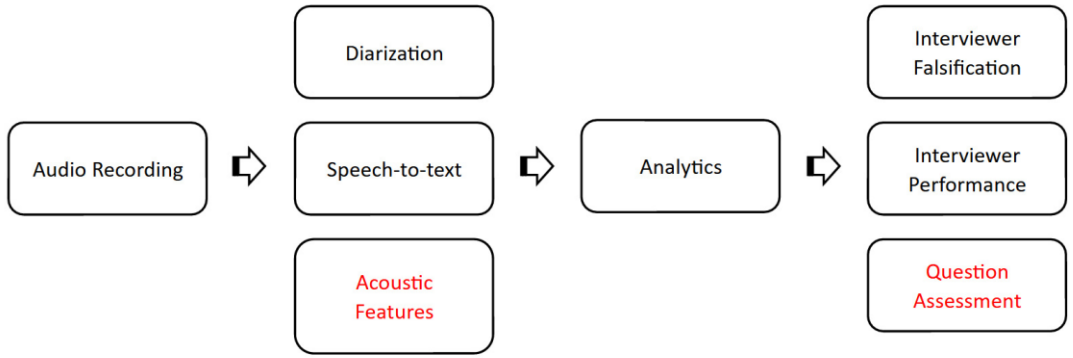
Scaling What We Already Know Works

NORC Research Science

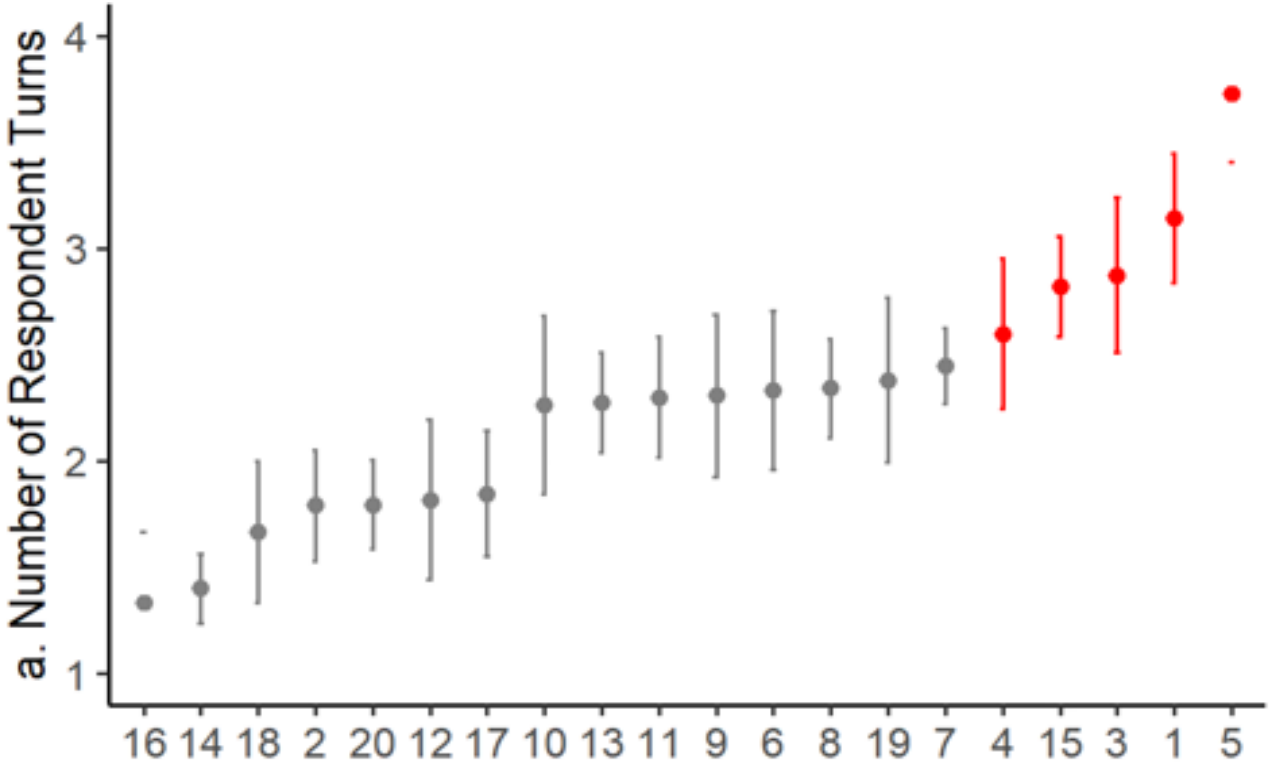
- Regular practice because of CARI
- Used together with behavior coding for
  - Interviewer behavior monitoring
  - Question assessment
- Human coding is labor intensive, time consuming, costly, and prone to unreliability
  - Only a small proportion of recordings are listened to and coded

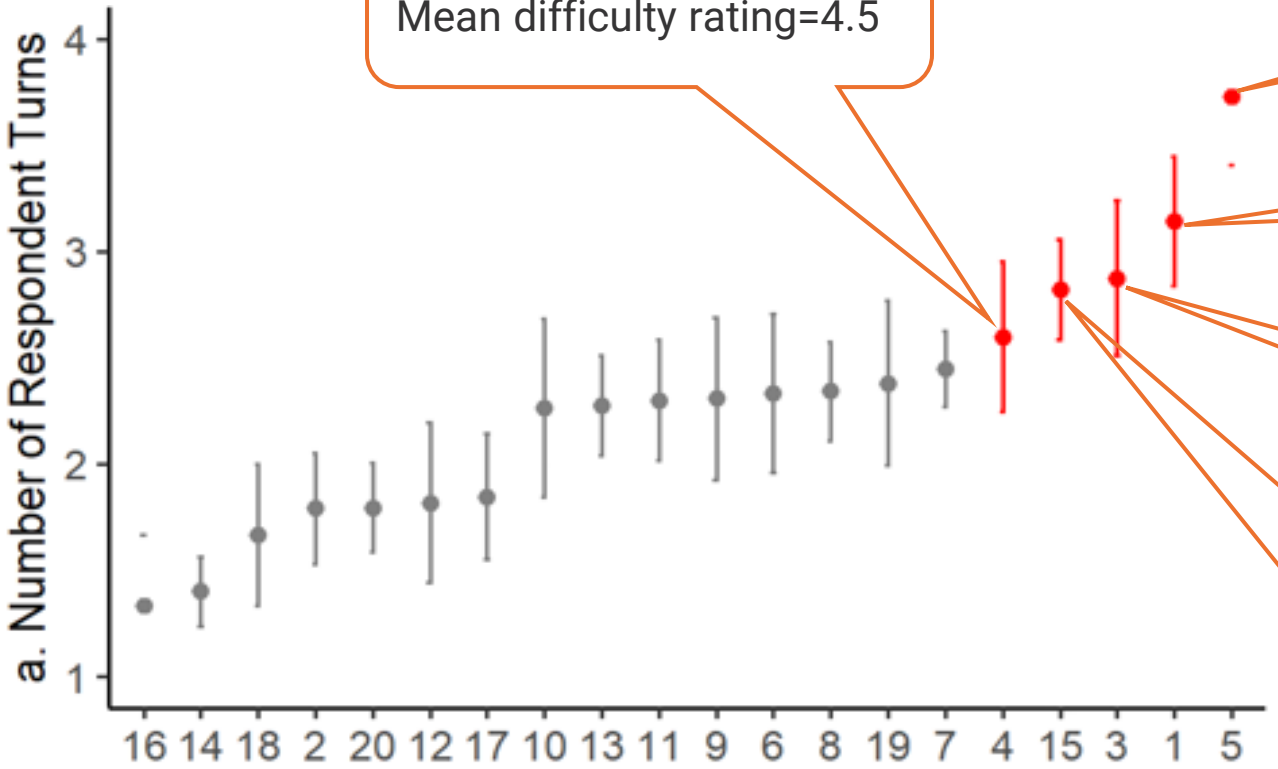


- # Machine Learning (ML) and Recordings of Interviews
  - **ML improves efficiency and speed of manually listening and coding recordings**
    - 100% of recordings processed automatically with minimal human involvement
  - **Measures indicative of poor interviewer behaviors**
  - **Measures indicative of poor performance, derived from paradigmatic question-answer sequence**



(Sun and Yan, 2023; Yan et al., 2024, 2025)


a. Number of Respondent Turns
4
3
2
1
16 14 18 2 20 12 17 10 13 11 9 6 8 19 7 4 15 3 1 5



Flagged 5 times;
Mean difficulty rating=4.5

Flagged 3 times;
Mean difficulty rating=3

Flagged 2 times
Mean difficulty rating=4.5

Flagged 4 times;
Mean difficulty rating=4.5

Flagged 3 times;
Mean difficulty rating=3

Why this is a responsible use of AI?

Intentional

Solving a bottleneck issue

Fit-for-purpose

An effective, efficient, and flexible triage tool in production

Human judgment and accountability

Survey methodologists led the design and evaluations and made key decisions

TSE-guided validation and Evaluation

Multifaceted validations were conducted along the way and/or built into the system

Transparent

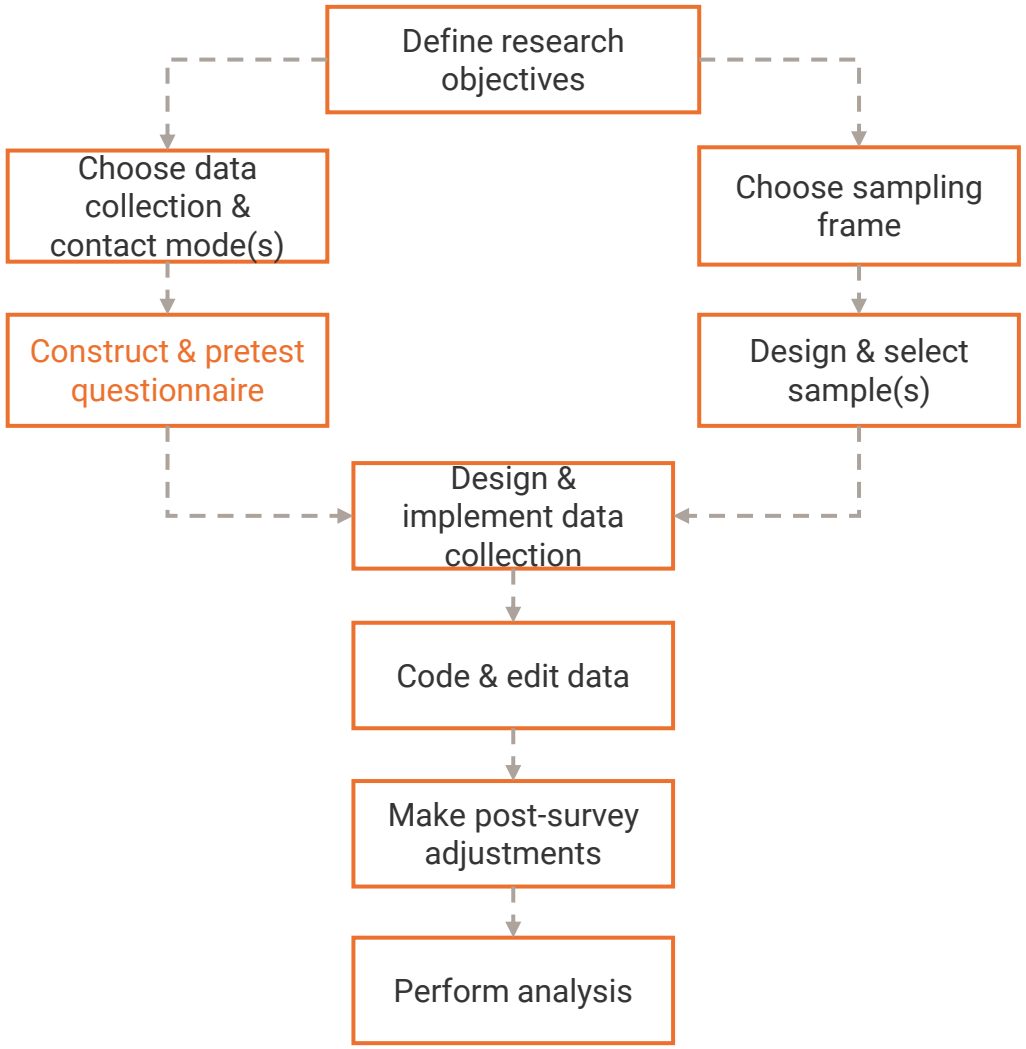
Documentation for understanding and reproducibility if possible

Example 2: Use of AI to Assist Questionnaire Evaluation

Developing Workflow for Production

NORC Research Science

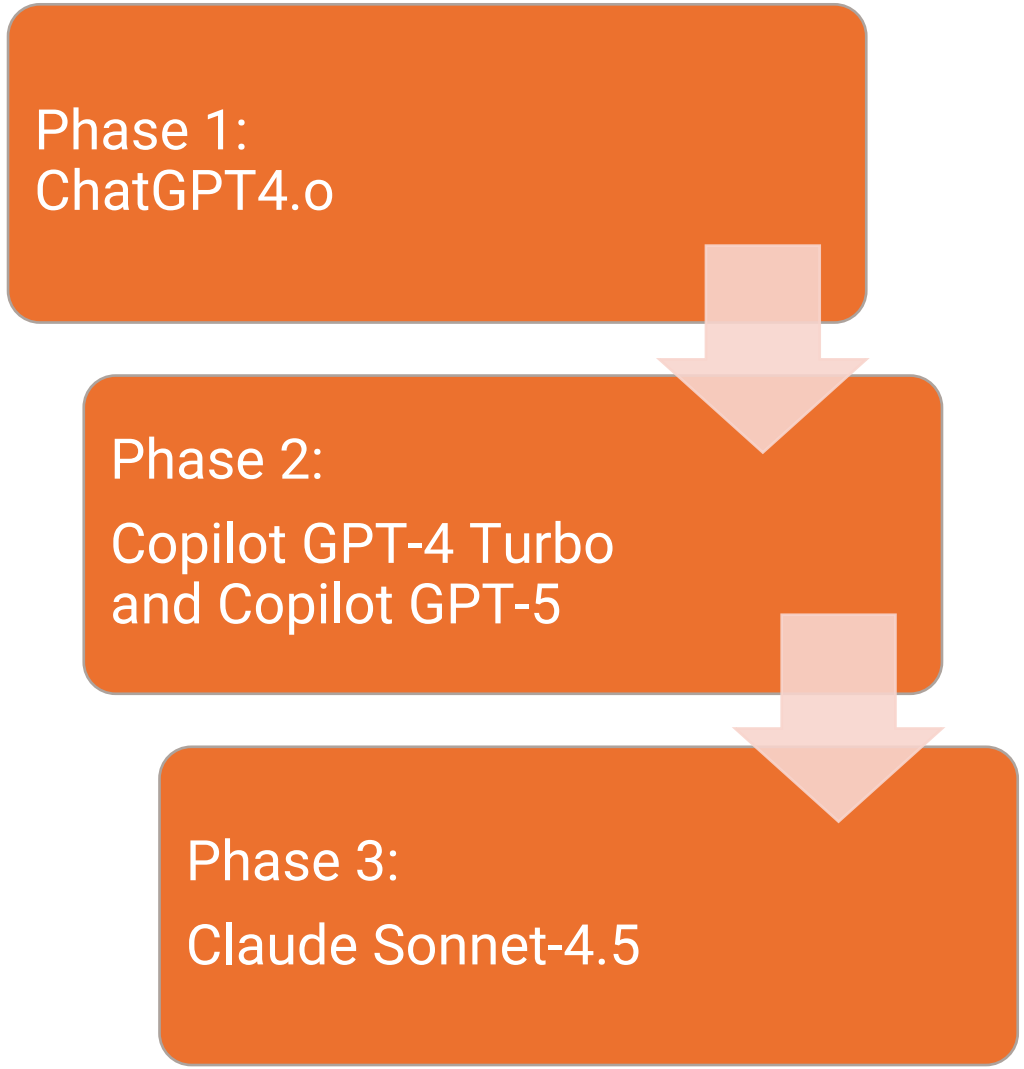
- A wide variety of question evaluation methods available that differ in cost, timeline, personnel, and type of information yielded
- Can be expensive and time-consuming, especially when participants are involved (e.g., cognitive interviews, field test)
- Expert reviews are considered less expensive and faster than lab-based and field-based methods



- Using AI to assist structure expert reviews
- Three phases
  - Same set of questions used in all phases

- Expert reviews from two survey methodologists
- One non-expert conducted review using QAS
- More results will be presented at AAPOR (2:15pm on Thursday May 14, 2026)

  https://aapor.confex.com/aapor/2026/meetingapp.cgi/Paper/5544

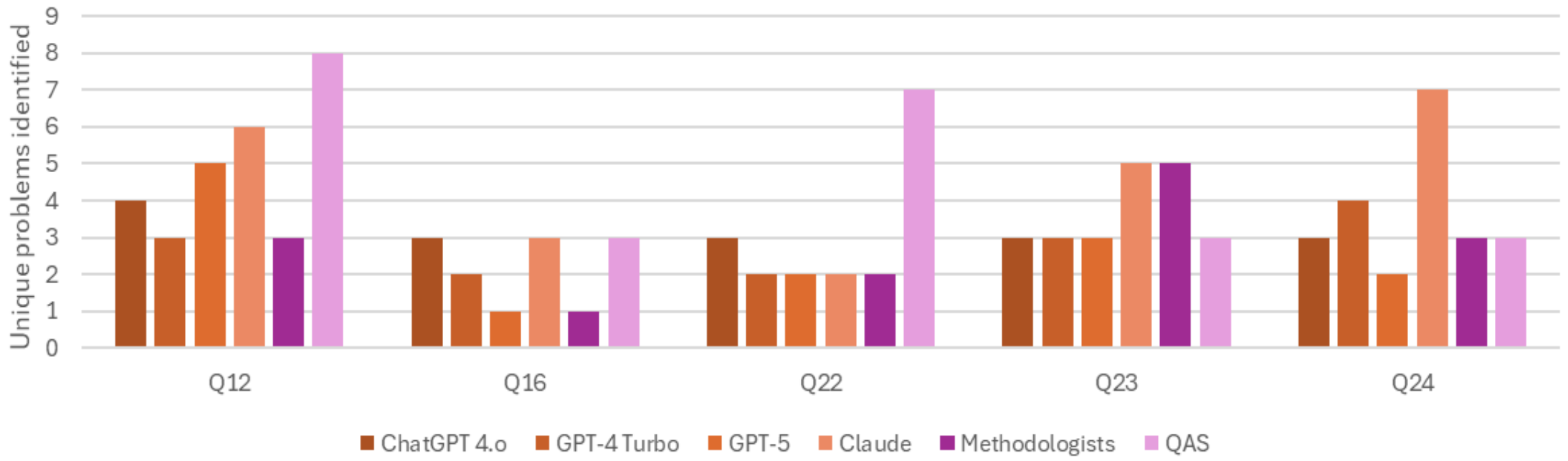


LLMs produced different number of unique problems

Claude > ChatGPT 4.o > GPT-4 T > GPT 5

QAS found most issues

Methodologists found fewer problems than QAS and most of LLMs



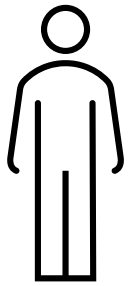
NORC (2026)

Q22

Enter any number between 0 and 100 that describes your quality of life:

1. Enter any number [Allow 0-100]
2. Prefer not to answer

Response Categories



Mapping the abstract concept to a numeric scale 0-100 may be difficult without more guidance/labels

Asking for a numeric rating (0-100) with no guidance can make it hard for respondents to map their complex lived experience to a number.

Claude

Lacks anchor definitions for what 0 and 100 represent, leading to different interpretation

ChatGPT

0-100 is likely to be too broad for R to come up with a meaningful number

Numeric scale may be hard to interpret without examples.

GPT 4 Turbo

GPT-5

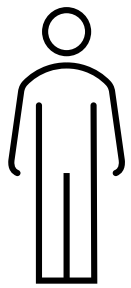
Numeric self-rating may be confusing

Q22

Enter any number between 0 and 100 that describes your quality of life:

1. Enter any number [Allow 0-100]
2. Prefer not to answer

Order effects



Order effects – previous questions will influence interpretation of “quality of life”.

GPT 4 Turbo

GPT-5

Claude

ChatGPT

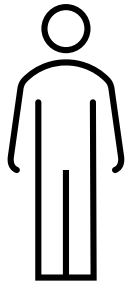
NORC (2026)

Q22

Enter any number between 0 and 100 that describes your quality of life:

1. Enter any number [Allow 0-100]
2. Prefer not to answer

Clarity



GPT 4 Turbo

Claude

"Quality of life" is not defined.

No explanation of what "quality of life" means; highly subjective.

GPT-5

"Quality of life" is a broad and subjective concept. Respondents may interpret it differently—some may think of physical health, others of emotional well-being, financial security, etc.

ChatGPT

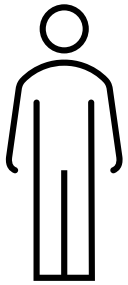
NORC (2026)

Q22

Enter any number between 0 and 100 that describes your quality of life:

1. Enter any number [Allow 0-100]
2. Prefer not to answer

Sensitivity



Claude

GPT-5

quality of life is highly personal and may feel intrusive

GPT 4 Turbo

People with low quality of life may feel vulnerable or uncomfortable quantifying their experience

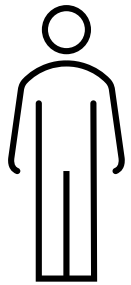
ChatGPT

Q22

Enter any number between 0 and 100 that describes your quality of life:

1. Enter any number [Allow 0-100]
2. Prefer not to answer

Burden/Motivation



GPT 4 Turbo

GPT-5

ChatGPT

Claude

Asking R for a number to a complex, multidimensional construct increases burden

NORC (2026)

>80% of problems identified by AI are valid concerns, but not all necessarily need a fix

AI good with structural and usability issues (e.g., long lists, lack of neutral options)

AI overreported sensitivity and stigma concerns, overemphasized burden issues, and missed nuanced comprehension issues, and cultural framing and implicit assumption issues

Claude outperformed other LLMs

- **Recommended incorporation into workflow as a first-pass triage tool**

Why this is a responsible use of AI?

Intentional

Improving efficiency

Fit-for-purpose

First-pass triage to flag problematic questions for further review

Human judgment and accountability

Survey methodologists led the design and made key decisions

TSE-guided validation and Evaluation

Validations conducted along the way and by phase

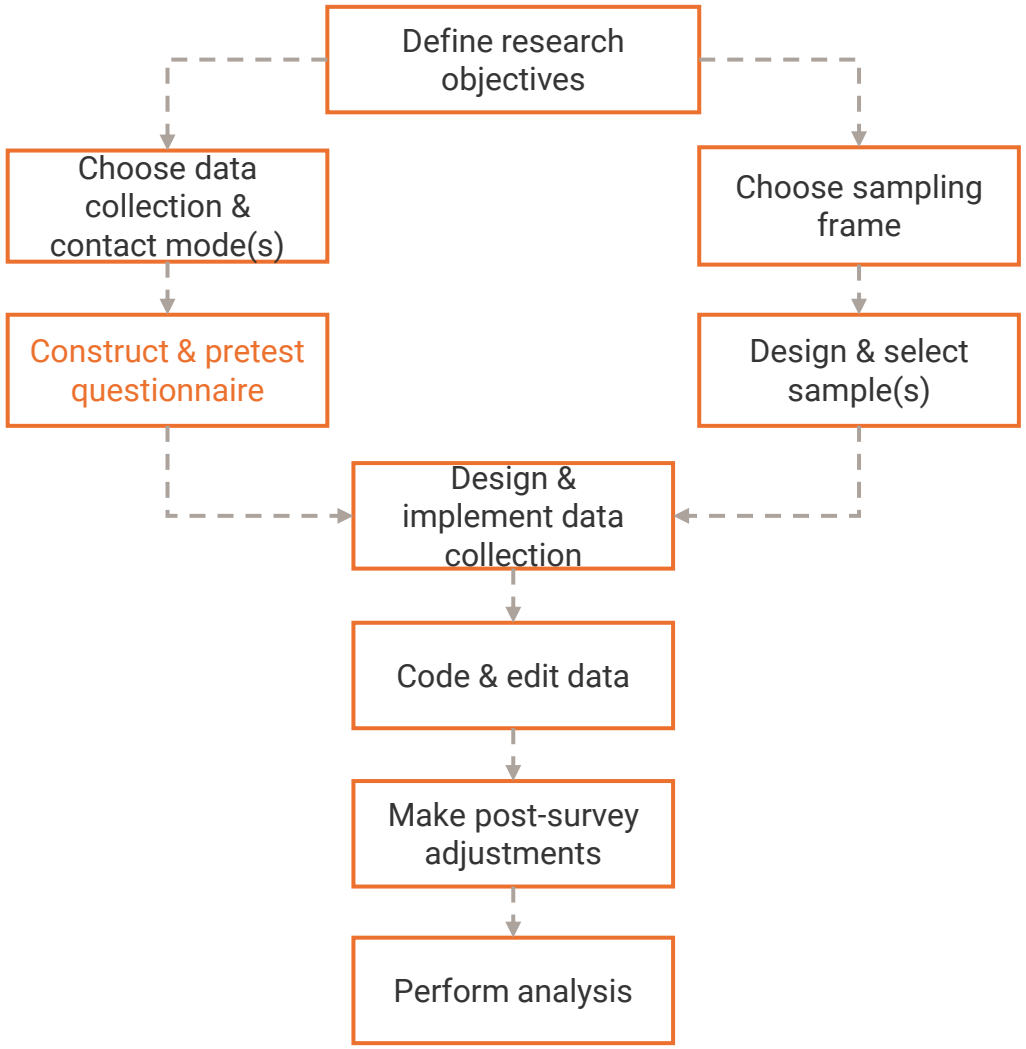
Transparent

Documentation for understanding and reproducibility if possible

Example 3: Use of AI to Test Programmed Web Surveys

NORC Research Science

- Web Survey Testing
  - One additional, important layer is the testing of the programmed instrument to check for programming errors in skip logic, out-of-range checks, typos etc
  - Labor intensive, time-consuming, and impossible to test all possible paths



A proof-of-concept study testing feasibility of using Claude to test web surveys

Short web survey programmed via Qualtrics

- Two skip logic errors
  - "Under 18" is allowed to proceed whereas any option over 18 is sent to a question asking for an adult
  - Answering "no" to two filter questions is shown a follow-up question
- One typo
  - "Forks" (instead of yes)
- One validation error
  - Only allowed to enter up to 3 to question about number of children

More results at AAPOR (10:15am on Friday May 15)

https://aapor.confex.com/aapor/2026/meetingapp.cgi/Paper/5460

Asked Claude to test survey as *a mother of 4 children*

Planned

- Two skip logic errors
  - "Under 18" is allowed to proceed whereas any option over 18 is sent to a question asking for an adult
  - Answering "no" to two filter questions is shown a follow-up question
- One typo
  - "Forks" (instead of yes)
- One validation error
  - Only allowed to enter up to 3 to question about number of children

Claude's feedback

-When selecting the 25-34 age range, the survey incorrectly displayed a message suggesting the respondent was under 18.... The survey should show the "under 18" message for respondents who select that age range

-After answering "No" to providing childcare for others' children, the survey still asks where children are being looked after... The flow doesn't properly skip irrelevant questions based on previous answers

-Typo in question S2c where "Forks" appears as an option instead of "Yes"

-Maximum children limit of 3 is too restrictive and doesn't reflect real-world family sizes

NORC (2025)

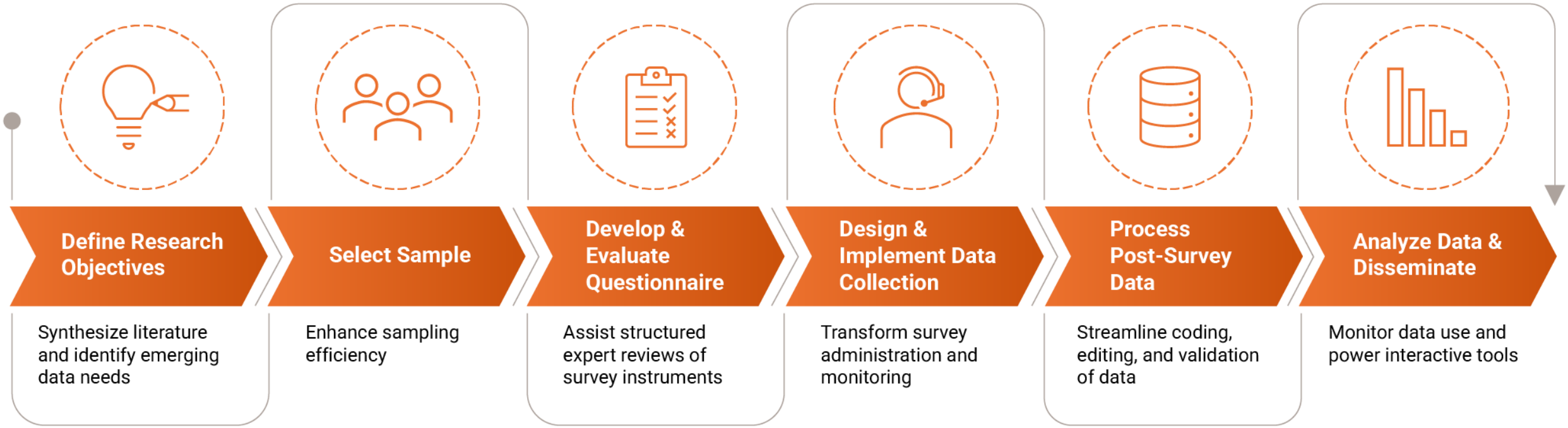
- Not fit for production
- We are moving away from this approach to a new framework that allows us to evaluate web instruments against its source, among other things
  - Wong et al. (2026): Tuesday April 21 at 4:30pm

NORC (2025)

Concluding Thoughts

NORC Research Science

Survey Methodologists Integral to Process of Embedding AI throughout Survey Life Cycle to Improve Efficiency and Data Quality



What about our respondents?

Trust

Transparency

Consent

Survey researchers are needed!

AI does not replace survey methodologists!

AI lacks:

- **Theoretical frameworks on errors, quality**
- **Design intent**
- **Accountability**
- **Ethical judgement**

Survey Methodologist provide:

- **Error-cost tradeoff reasoning**
- **Validity checks**
- **Human subject protections**

Future of survey research is methodologist-driven with AI in service of our goals!

Ting Yan
yan-ting@norc.org

Research You Can Trust™

NORC Research Science

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 Annual Conference of American Association for Public Opinion Research. Accessed at: https://www.norc.org/content/dam/norc-org/pdf2025/Barari%20--%20AAPOR%20presentation.pdf

Buskirk, T. D., Keusch, F., von der Heyde, L., & Eck, A. (2025). 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

Rothschild, D.M., Buskirk, T.D., Eckman, S., Hillygus, D.S., Kreuter, F., and Lazer, D. (2025). New and Emerging Methods: Successfully Navigating the Disruption AI will Bring to Survey Research. The Survey Statistician, 92, 30-44. https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2025_July_N92_04.pdf.

Buskirk, T. D. (2026). Let's Not Leave Probability Panels to Chance: Why AI Matters for Their Future. Keynote speech at the 2026 CIPHER Conference. Accessed at: https://dornsife.usc.edu/cesr/wp-content/uploads/sites/54/2026/03/CIPHER-2026-Buskirk-Keynote.pdf.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M. Singer, E., and Tourangeau, R. (2009). *Survey Methodology*. John Wiley & Sons Inc., Hoboken.

Sun, H., and Yan, T. (2023). Applying Machine Learning to the Evaluation of Interviewer Performance. *Survey Practice*, 16 (1). https://doi.org/10.29115/SP-2023-0007.

Yan, T., Sun, H., and Battalahalli, A. (2024). Applying Machine Learning to Survey Question Assessment. *Survey Practice*, https://doi.org/10.29115/SP-2024-0006.

Yan, T., Sun, H., and Battalahalli, A. (2025). Using Machine Learning to Evaluate Questions in a Multilingual Survey. *Survey Practice*, 19 Special Issue (March). https://doi.org/10.29115/SP-2024-0021.

NORC (2026). Use of Large Language Models to Assist Expert Reviews. Internal Memo.

NORC (2025). Using Claude to Test Web Survey Instrument. Internal Memo.