SURVEYS · TECHNOLOGY AND THE FUTURE

Artificial Intelligence in Questionnaire Design and Analysis

Artificial intelligence is now used to draft survey questions, test them, code open-ended answers and model survey data. This article explains where AI fits in the survey process, what the evidence shows about its accuracy, the risks it brings and how researchers can use it responsibly.

ModQu research article · 2,751 words · 26 references (APA 7th edition)

Artificial intelligence (AI) has moved quickly from a specialist research topic into everyday survey work. Researchers now use AI tools to suggest question wordings, check questionnaires for problems, translate them, sort thousands of open-ended answers into categories and build statistical models from survey data. Some have gone further and asked whether AI can stand in for human respondents altogether.

These developments raise practical questions. Which tasks can AI do well, and which still need human judgement? How accurate is it compared with trained researchers and coders? What new risks does it introduce, and how should its use be documented? This article reviews the main uses of AI across the survey process, summarises what research has found so far and sets out principles for using it responsibly.

What "artificial intelligence" means here

The term covers several different technologies. **Machine learning** refers to statistical methods that learn patterns from data, such as decision trees and random forests that predict an outcome from many variables. Survey researchers have used these methods for years to predict nonresponse, build weights and classify text (Buskirk et al., 2018; Kern et al., 2019).

Natural language processing is the branch of AI that works with text. It includes earlier methods that count words and phrases, such as topic models, and newer methods based on neural networks (Grimmer & Stewart, 2013).

Large language models are neural networks trained on very large amounts of text to predict the next word in a sequence. Because they can produce fluent text and follow instructions written in plain language, they can be applied to many survey tasks without specialist programming. Most current interest in AI for surveys concerns these models. They are also the source of most of the new risks, because their output can be fluent and confident even when it is wrong (Ji et al., 2023).

Figure 1 shows where these tools can be used at each stage of a survey.

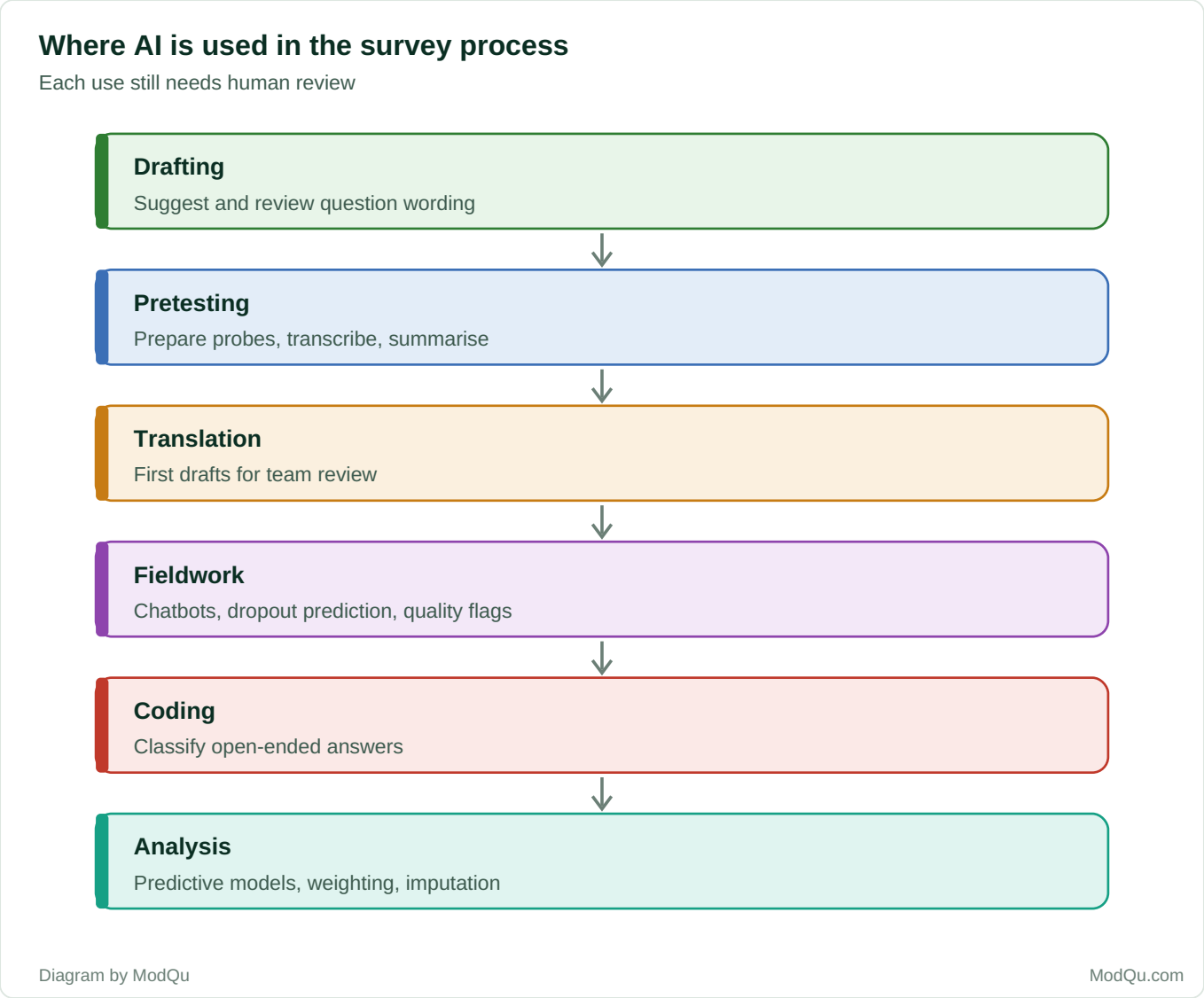


Figure 1. Uses of AI at each stage of a survey. Diagram by ModQu.

Earlier computer aids to questionnaire design

Computer tools for questionnaire design are not new. In the early 2000s, Graesser et al. (2006) developed QUAID, a web tool that flagged features of questions likely to cause comprehension problems, such as unfamiliar technical terms, vague or imprecise wording, complex syntax and heavy demands on working memory. The tool drew on research showing that such features increase the effort respondents need to understand a question. Lenzner et al. (2010) later showed experimentally that questions with these features took longer to answer, a sign of greater cognitive burden.

Another tool, the Survey Quality Predictor, uses results from many experiments to estimate the measurement quality of a question from its characteristics, such as the number of response categories, the use of labels and the presence of a "don't know" option (Saris & Gallhofer, 2014). These tools did not write questions. They applied rules and evidence from survey methodology to help researchers spot weaknesses in their own drafts.

AI in drafting and reviewing questions

Large language models can now produce draft questions from a short description of a topic. A researcher might ask a model to suggest ten questions about attitudes to remote working, or to rewrite a question in plainer language. Models can also review a draft against a checklist of common problems: double-barrelled questions that ask about two things at once, leading wording, unbalanced response scales, undefined terms and missing response options.

Used in this way, AI acts as a fast, tireless first reviewer. It can catch problems that a busy researcher might miss and suggest alternatives. But it does not know the purpose of the survey, the population being surveyed or the concepts the researcher intends to measure unless it is told, and it has no direct experience of how real respondents understand words. A question that looks clear to a model trained on written text may still be misunderstood by respondents with different backgrounds. Draft questions from AI therefore need the same scrutiny as any other draft.

The established theory of how people answer questions, which describes comprehension, retrieval, judgement and response as separate steps, remains the right framework for judging any draft, whatever its source (Tourangeau et al., 2000). AI suggestions should be judged on whether they make each of these steps easier for respondents.

AI in pretesting

Pretesting is the stage at which questions are tried out before the main survey. Established methods include expert review, cognitive interviewing and pilot studies (Presser et al., 2004). In cognitive interviews, a researcher asks a small number of people to answer the questions and to describe how they arrived at their answers. These interviews reveal misunderstandings that experts often miss (Willis, 2005; Beatty & Willis, 2007).

AI can support pretesting in several ways. It can help prepare interview guides and probes, transcribe recorded interviews and summarise patterns across transcripts. Some researchers have also asked language models to answer draft questions and explain their reasoning, as a quick screen for ambiguity before human testing.

This last use has clear limits. Research on using language models to simulate survey respondents shows that they can reproduce some broad patterns in public opinion but also give answers that are too uniform, that vary with small changes in the prompt and that misrepresent some groups (Argyle et al., 2023; Bisbee et al., 2024). A model's answers are not evidence of how real people will understand a question. AI can make pretesting faster and cheaper, but it cannot replace testing with members of the target population.

AI in translation

Surveys conducted in several languages need translations that preserve meaning, not just words. AI translation tools now produce fluent translations quickly. They can serve as one of the initial translations in a team process, which is then reviewed and adjusted by human translators who know the survey's concepts and the cultures involved. Using an unchecked machine translation as the final version risks errors that change the meaning of questions and make comparisons between languages invalid.

AI during fieldwork

During data collection, AI can make surveys more responsive. **Chatbot surveys** use conversational software to ask questions and follow up on answers. Xiao et al. (2020) compared a chatbot that used AI to respond to open-ended answers with a conventional online survey and found that the chatbot drew longer, more informative and more relevant answers. Conversational surveys are discussed in a separate article in this series.

Machine learning is also used to manage fieldwork. Models trained on earlier waves of a panel or on paradata can predict which sample members are likely to drop out, so that extra effort can be directed to them. Kern et al. (2023) found that machine learning methods predicted nonresponse in future waves of a panel reasonably well, which could support targeted strategies to keep people in the panel. Similar models can flag likely careless responses during data collection so that they can be reviewed.

AI in checking data quality

Online surveys face growing problems with careless respondents and with fraudulent answers, including answers written by automated programs. Machine learning offers ways to detect them. Instead of applying a single rule, such as removing everyone who answered faster than a fixed time, a model can combine many signals: response times, patterns in grid answers, inconsistent answers, the text of open-ended answers and device information. The model learns which combinations distinguish genuine from problematic responses in data where the answer is known.

This approach brings its own difficulties. Models trained on one survey may not work on another, and a model that removes responses wrongly can bias results against groups who answer differently for legitimate reasons, such as respondents who read slowly or answer on older devices. Flags produced by a model should therefore lead to review, not to automatic deletion. Language models also create a new problem: they make it easy to produce fluent, plausible open-ended answers at scale, so that text quality alone is no longer a reliable sign of a genuine respondent. Detection methods will need to keep adapting.

AI in coding open-ended answers

Open-ended questions let respondents answer in their own words. They are valuable but expensive to analyse, because each answer must be read and assigned to a category, a task called **coding**. AI has had perhaps its greatest effect here.

Supervised machine learning methods learn from a set of answers already coded by people and then code the rest. Schonlau and Couper (2016) proposed a semi-automatic approach in which the computer codes the answers it can classify with high confidence and passes the rest to human coders. This saves much of the effort while keeping accuracy high for difficult cases. Similar methods have been developed for coding occupations, a large and standard task in official statistics (Gweon et al., 2017; Schierholz & Schonlau, 2021).

Unsupervised methods find themes without a prior coding frame. **Structural topic models**, for example, identify groups of words that tend to occur together and relate them to respondent characteristics (Roberts et al., 2014).

Large language models can code answers from instructions alone, with little or no training data. Gilardi et al. (2023) found that a language model outperformed crowd workers on several text annotation tasks and cost far less. Mellon et al. (2024) used language models to code open-ended answers to a question about the most important issue facing the country in a large British survey and found that their accuracy was close to that of trained human coders. A review of language models in social science research found that their performance varied widely between tasks, being strong for some classification tasks and weak for others, so that each use needs its own validation (Ziems et al., 2024).

Figure 2 illustrates how agreement with expert coders might compare across methods in a validation study.

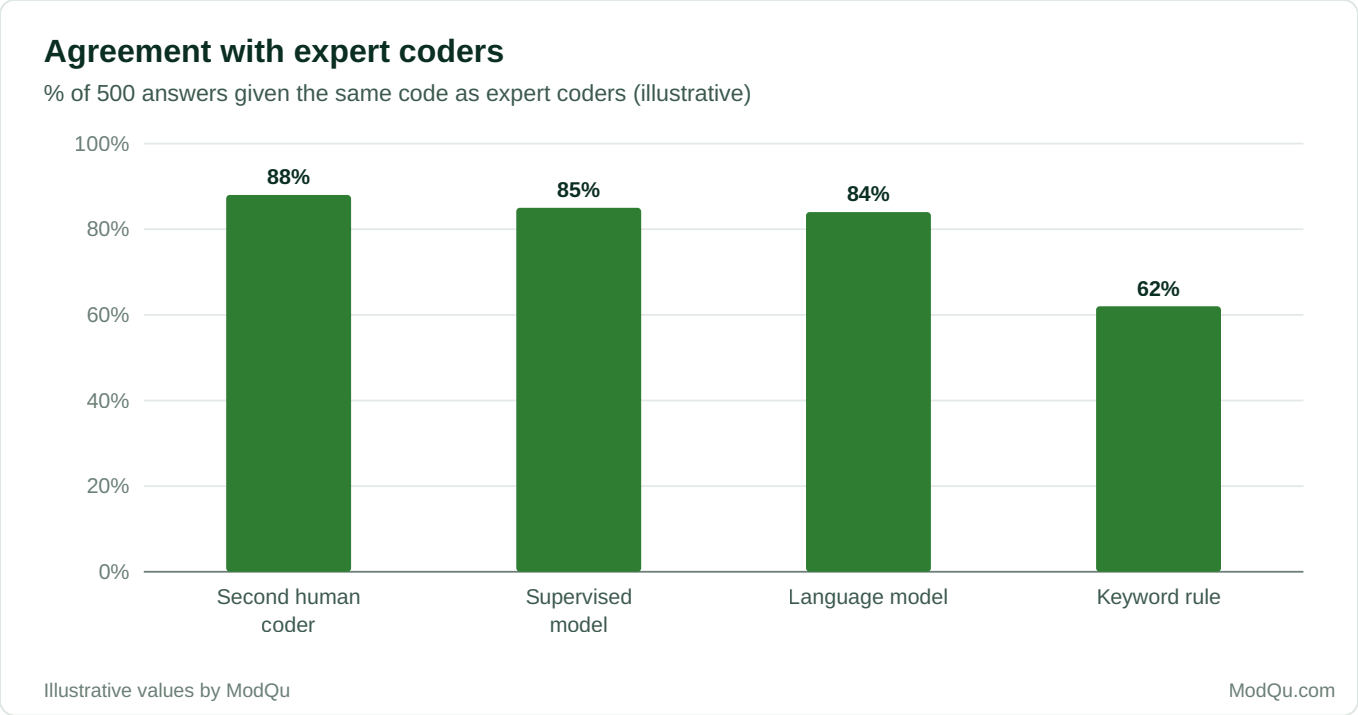


Figure 2. Agreement with expert coders for different coding methods in a hypothetical validation study (illustrative values). Chart by ModQu.

AI in analysis

Machine learning methods are now part of many survey analyses. Tree-based methods, such as random forests, can model complex relationships without the analyst specifying them in advance. They are used to model response propensities for weighting, to impute missing values and to explore which variables best predict an outcome (Buskirk & Kolenikov, 2015; Kern et al., 2019). Buskirk et al. (2018) give an introduction to these methods for survey researchers, emphasising that predictive accuracy is not the same as explanation and that survey design features, such as weights and clustering, still need attention.

Language models are also used to write code for analysis, to summarise results and to draft reports. Each of these uses saves time, but each can introduce errors that are hard to see. A summary may describe a pattern that is not in the data, and generated code may run without error yet compute the wrong statistic. Analysts should check generated code line by line and confirm every figure in a generated summary against the data.

Risks and limitations

Errors presented with confidence

Language models produce text that reads well, whether or not it is correct. They can invent facts, references and statistics, a problem known as **hallucination** (Ji et al., 2023). In survey work, this can mean a summary that misstates a result or a literature review that cites studies that do not exist. Every factual claim produced by a model needs checking against its source.

Bias

Models learn from the text on which they are trained, and that text over-represents some groups, languages and viewpoints. Santurkar et al. (2023) compared the opinions expressed by several language models with those of different groups in the United States and found that models' default answers were closer to some groups than to others. Bender et al. (2021) warned that very large language models can reproduce and amplify social biases present in their training data. When AI codes answers or drafts questions, these biases can enter survey results without anyone noticing.

Reproducibility

Commercial models change over time, and the same instruction can produce different answers on different days. A coding exercise carried out with one version of a model may not be reproducible a year later. Researchers should record the model, version, date and full instructions used, and keep the outputs, so that others can see exactly what was done.

Privacy

Survey answers often contain personal or sensitive information. Sending them to an external AI service may conflict with the promises made to respondents and with data protection law. Before using AI on survey data, researchers should check where the data will be processed and stored, whether the provider uses it to train further models and whether respondents' consent covers this use. Removing names and other identifiers before processing reduces the risk but does not remove it.

Loss of expertise

If routine tasks are handed to AI, researchers may lose the experience that comes from reading answers and wrestling with question wording. That experience is how researchers learn what respondents mean. AI should support this learning, not replace it.

Skills researchers will need

Using AI well requires a mix of old and new skills. Researchers still need a sound understanding of question design, sampling and measurement, because these are the standards against which AI output must be judged. They also need enough understanding of how AI tools work to recognise their typical failures: a language model's tendency to produce plausible text regardless of truth, a classifier's tendency to perform poorly on categories that were rare in its training data and a predictive model's tendency to fit patterns that do not hold in new data.

Organisations that use AI in survey work benefit from written procedures that say which tasks AI may be used for, what validation is required and who signs off the results. Such procedures make good practice routine rather than dependent on individual judgement, and they make it easier to explain the methods to clients and readers.

Principles for responsible use

Several principles follow from the evidence so far:

- **Keep a human responsible.** A named researcher should review and take responsibility for every question, code and conclusion, whatever tools were used.
- **Validate against human judgement.** Before using AI to code open-ended answers, compare its codes with those of trained coders on a sample, and report the level of agreement. Repeat the check when the model or the question changes.
- **Test with real people.** Use AI to improve drafts, not to replace cognitive interviews and pilots with members of the target population.
- **Document fully.** Record which tools were used, for which tasks, with which settings and instructions. Professional codes already require researchers to disclose their methods, and AI use is part of the method (American Association for Public Opinion Research, 2021).
- **Protect respondents.** Check data protection obligations before sending survey data to any AI service, and remove identifiers.
- **Check every fact.** Treat any statement, figure or reference generated by AI as unverified until checked.

An example

A market research team runs a customer survey with 6,000 answers to the question "What one thing would most improve our service?" In the past, two coders took about a week to code such answers. This time, the team writes a coding frame of 14 categories and asks a language model to assign each answer to a category.

Before relying on the results, two experienced coders independently code a random sample of 500 answers. The model agrees with the expert codes in 84 per cent of cases, close to the 88 per cent agreement between the two human coders, but performs poorly on one category, answers about staff attitude, which it often confuses with waiting times. The team revises the category definitions, reruns the model and has a coder review all answers the model classified in those two categories. The final coding takes two days instead of a week, and the report describes the method and the validation results (illustrative example).

Conclusion

AI offers survey researchers real gains in speed and scale, especially for coding open-ended answers, reviewing drafts and building predictive models. The evidence so far shows that it can match human performance on some tasks while failing badly on others, and that its errors can be hard to see. The value of AI in survey research depends on the methods that have always protected survey quality: careful question design, testing with real respondents, validation against human judgement and full disclosure of how the results were produced. Used within those methods, AI can help generate knowledge more efficiently. Used in place of them, it risks producing results that look authoritative but cannot be trusted.

References

American Association for Public Opinion Research. (2021). *AAPOR code of professional ethics and practices*. AAPOR. https://aapor.org/standards-and-ethics/

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & 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

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*(2), 287–311. https://doi.org/10.1093/poq/nfm006

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, *32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

Buskirk, T. D., & Kolenikov, S. (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2015-00003

Buskirk, T. D., Kirchner, A., Eck, A., & Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, *11*(1). https://doi.org/10.29115/SP-2018-0004

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

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*(1), 3–22. https://doi.org/10.1093/poq/nfj012

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics, 33*(1), 101–122. https://doi.org/10.1515/jos-2017-0006

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods, 13*(1), 73–93. https://doi.org/10.18148/srm/2019.v13i1.7395

Kern, C., Weiß, B., & Kolb, J.-P. (2023). Predicting nonresponse in future waves of a probability-based mixed-mode panel with machine learning. *Journal of Survey Statistics and Methodology, 11*(1), 100–123. https://doi.org/10.1093/jssam/smab009

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7), 1003–1020. https://doi.org/10.1002/acp.1602

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics, 11*(1). https://doi.org/10.1177/20531680241231468

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*(1), 109–130. https://doi.org/10.1093/poq/nfh008

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning* (pp. 29971–30004). PMLR.

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research* (2nd ed.). Wiley. https://doi.org/10.1002/9781118634646

Schierholz, M., & Schonlau, M. (2021). Machine learning for occupation coding: A comparison study. *Journal of Survey Statistics and Methodology, 9*(5), 1013–1034. https://doi.org/10.1093/jssam/smaa023

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods, 10*(2), 143–152. https://doi.org/10.18148/srm/2016.v10i2.6213

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE. https://doi.org/10.4135/9781412983655

Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction, 27*(3), Article 15. https://doi.org/10.1145/3381804

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics, 50*(1), 237–291. https://doi.org/10.1162/coli_a_00502