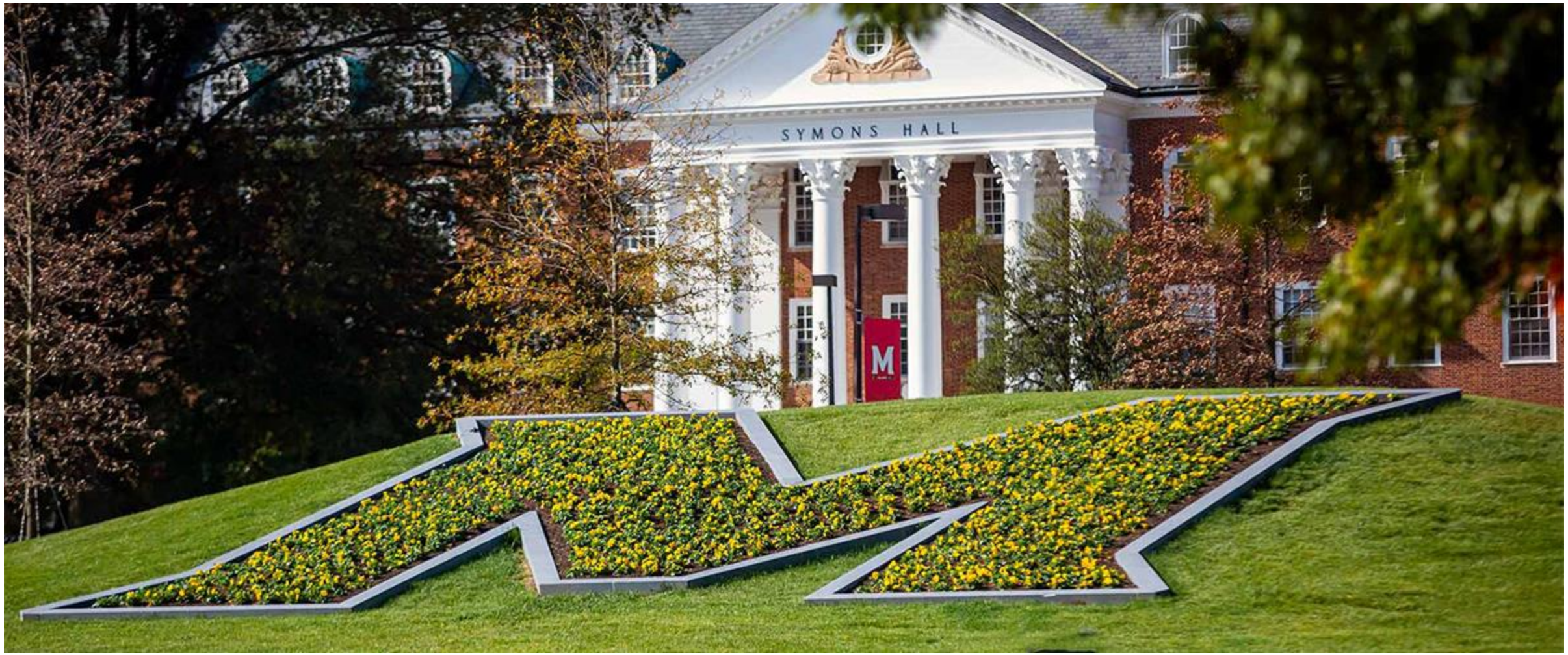



# Leveraging AI for Survey Statistics and Survey Research

Frauke Kreuter
JPSM and SoDa Center, University of Maryland & LMU Munich
Frauke.Kreuter@lmu.de

# Leveraging AI for Survey Research

QUESTIONNAIRE DESIGN

SYNTHETIC DATA

INTERVIEWER

# Leveraging Survey Research for AI

QUESTIONNAIRE DESIGN



SYNTHETIC DATA



INTERVIEWER

# #prompt including request to use Schuman and Presser

Needs prompts that include textbooks and good articles as context.

Given good prompting, certainly a valuable help.

Very good at translating to different reading levels.

1. **On a scale of 1-10, how satisfied were you with this class?** Revised: How would you describe your overall satisfaction with this class?
   Very dissatisfied, Dissatisfied, Neutral, Satisfied, Very satisfied.

**Explanation**: A 5-point scale simplifies the choices and captures nuanced feelings. Avoiding numbers ensures that participants don't just choose a 'middle' number without giving it thought.

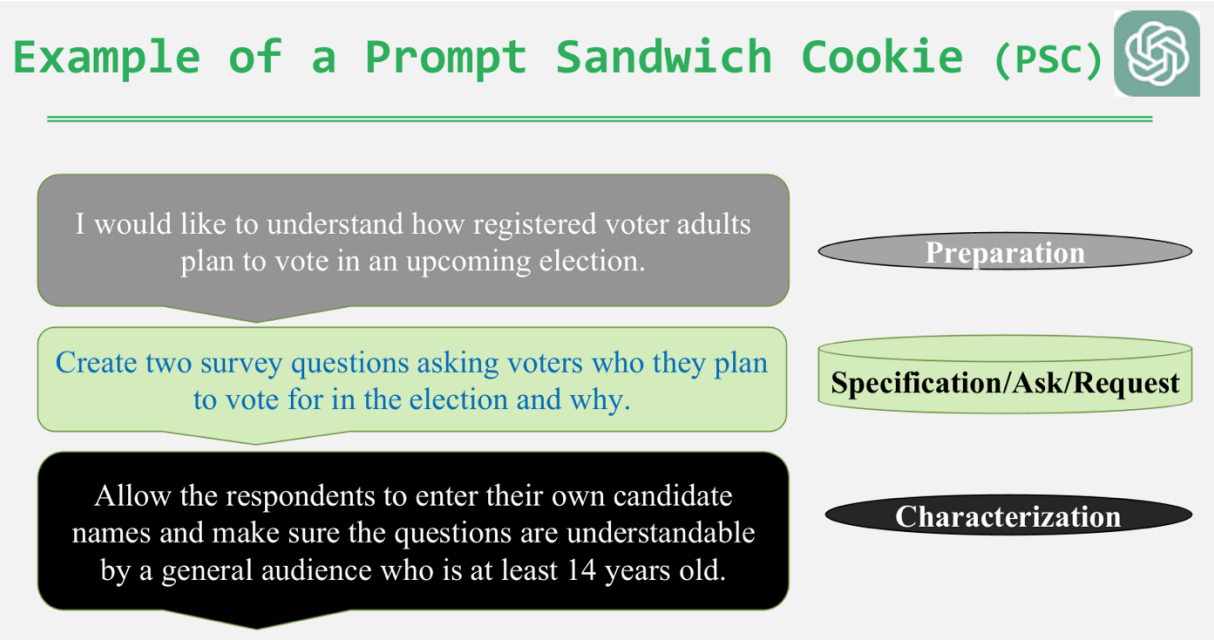
2. **How relevant and useful did you find the content of the class for your academic and/or professional goals?** (1 = Not relevant at all, 10 = Extremely relevant)
   Revised: How would you rate the relevance of the class content to your academic or professional goals? Options: Not at all relevant, Slightly relevant, Moderately relevant, Very relevant, Extremely relevant.

**Explanation**: Splitting relevance and usefulness makes it clearer and avoids confusion. Simplified options provide clarity. Not mentioned: Double Barrel….

Buskirk et al. (2024, AAPOR) designed a series of 4 sequential and empirical experiments aimed at learning how to create superprompts for LLMs to generate survey questions.

The experiments consider components and formats of prompts including:

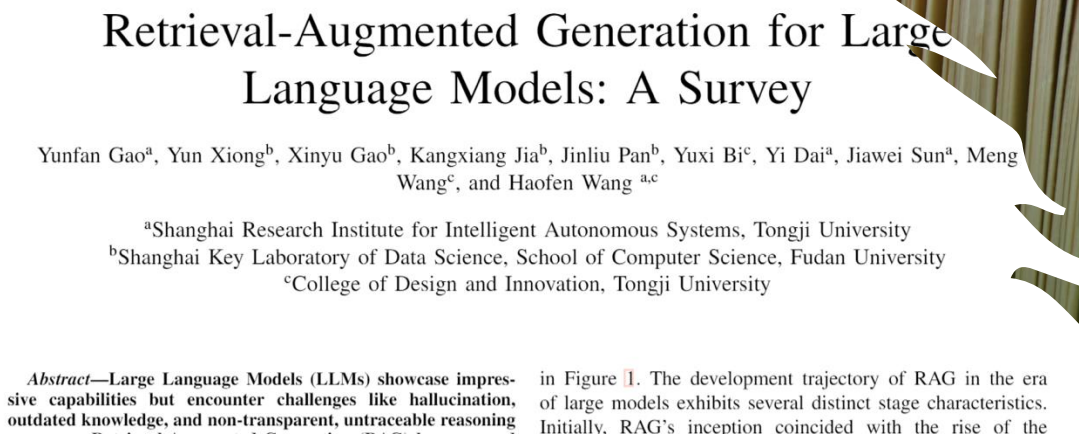
**Experiment 1**: Use of the keywords "survey" and "response options/answer choices"
**Experiment 2**: Complexity of the prompt to include requests for clarifications and parentheticals in the survey stem and responses
**Experiment 3**: Controlling the reading level of items/response options
**Experiment 4**: Controlling the content and number of response options that are generated.



Needs prompts that include textbooks and good articles as context.

Given good prompting, certainly a valuable help.

Very good at translating to different reading levels.

Retrieval-Augmented Generation for Large Language Models: A Survey

Yunfan Gao[a], Yun Xiong[b], Xinyu Gao[b], Kangxiang Jia[b], Jinliu Pan[b], Yuxi Bi[c], Yi Dai[a], Jiawei Sun[a], Meng Wang[c], and Haofen Wang [a,c]

[a]Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University
[b]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[c]College of Design and Innovation, Tongji University

*Abstract*—**Large Language Models (LLMs) showcase impressive capabilities but encounter challenges like hallucination, outdated knowledge, and non-transparent, untraceable reasoning**

in Figure 1. The development trajectory of RAG in the era of large models exhibits several distinct stage characteristics. Initially, RAG's inception coincided with the rise of the

QUESTIONNAIRE DESIGN

SYNTHETIC DATA



INTERVIEWER

**English (translation)** I am 28 years old and female. I have a college degree, a medium monthly net household income, and am working. I am not religious. Ideologically, I am leaning center-left. I rather weakly identify with the Green party. I live in West Germany. I think the government should facilitate immigration and take measures to reduce income disparities. Did I vote in the 2017 German parliamentary elections and if so, which party did I vote for? I [INSERT]

| | |
|---|---|
| | gious], sehr religiös [very religious] |
| leftright | stark links [strongly left], mittig links [center-left], in der Mitte [in the middle], mittig rechts [center-right], stark rechts [strongly right] |
| partyid_degree | sehr stark [very strongly], ziemlich stark [rather strongly], mäßig [moderately], ziemlich schwach [rather weakly], sehr schwach [very weakly] |
| partyid | mit der Partei CDU/CSU [CDU/CSU], mit der Partei SPD [SPD], mit der Partei Bündnis 90/Die Grünen [Greens], mit der Partei FDP [FDP], mit der Partei Die Linke [Left], mit der Partei AfD [AfD], mit einer Kleinpartei [small party], mit keiner Partei [not with any party] |
| east | 0 Westdeutschland [West Germany], 1 Ostdeutschland [East Germany] |
| immigration | erleichtern [facilitate], weder erleichtern noch einschränken [neither nor], einschränken [limit] |
| inequality | Maßnahmen ergreifen [take measures], habe keine Meinung dazu, ob die Regierung Maßnahmen ergreifen sollte [no opinion], keine Maßnahmen ergreifen [don't take measures] |


Political Analysis

## Article contents



# 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 | Supplementary materials | Metrics

Get access | Share | Cite | Rights & Permissions

## 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 applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), which are often treated as uniform properties of the models. We show that the "algorithmic bias" within one such tool—the GPT-3 language model—is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause it to accurately emulate response distributions from a wide variety of human subgroups. We term this property *algorithmic fidelity* and explore its extent in GPT-3. We create "silicon samples" by conditioning the model on thousands of sociodemographic backstories from real human participants in multiple large surveys conducted in the United States. We then compare the silicon and human samples to demonstrate that the

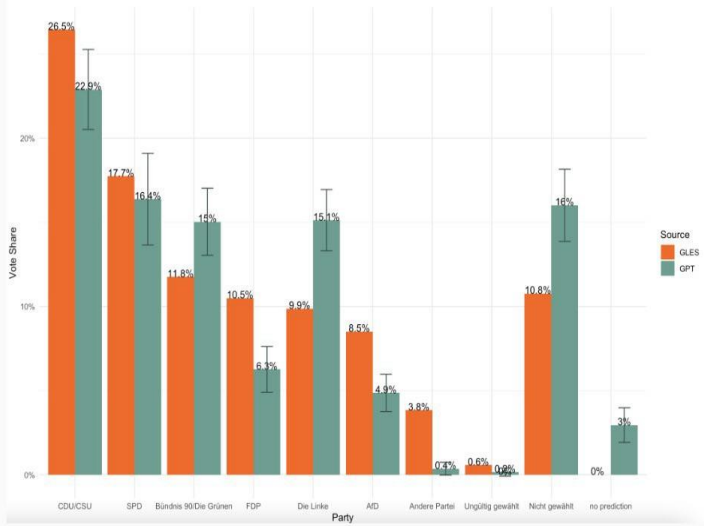
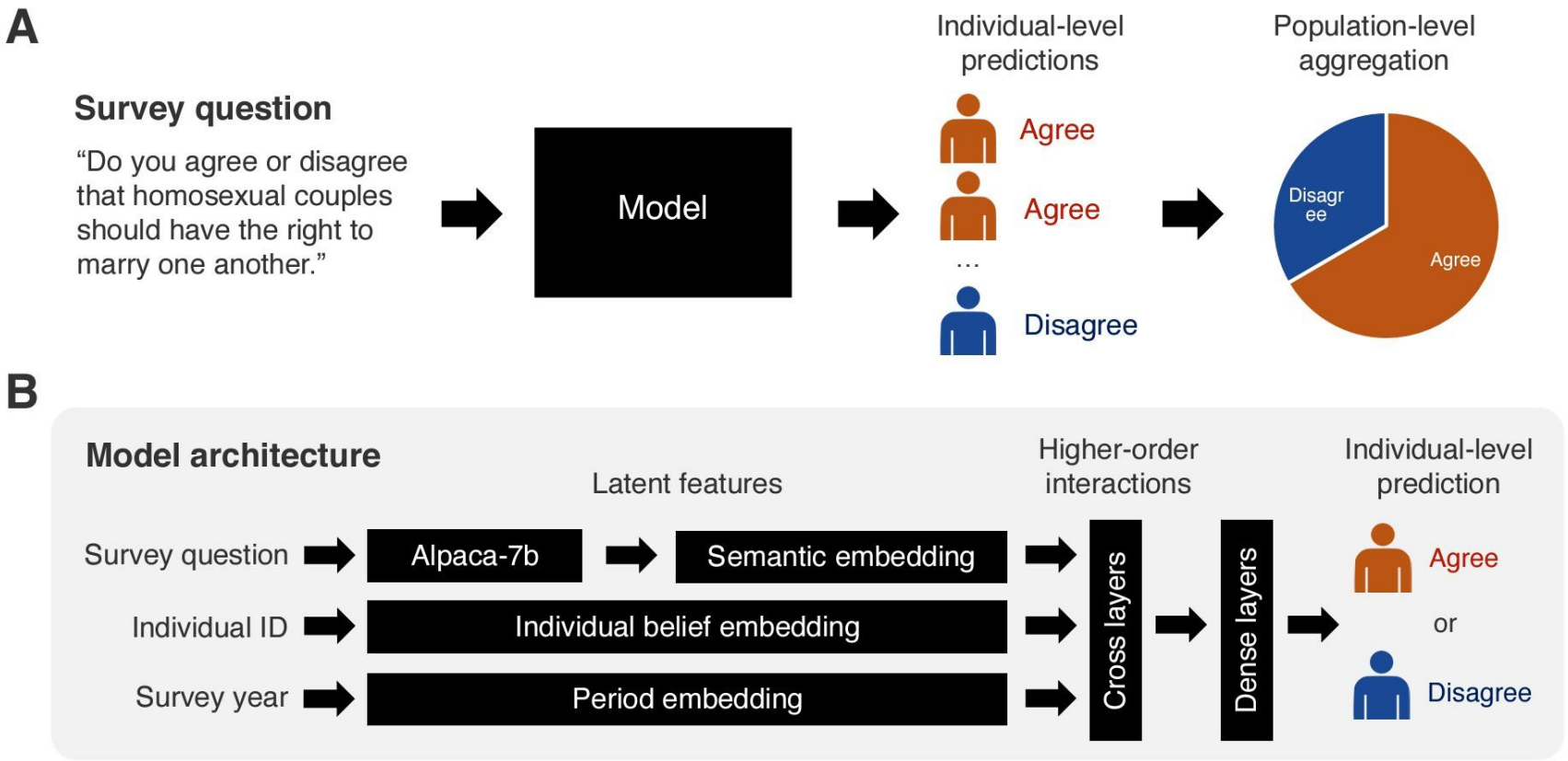



**Figure 3:** Replicating Argyle et al. for German data (GLES): Current project by Leah von der Heyde, Alexander Wenz and Carolina Haensch



Von der Heyde, L., Wenz, A., & Haensch, A.-C. (2024, February 22). Artificial Intelligence, Unbiased Opinions? Assessing GPT's suitability for estimating public opinion in multi-party systems. https://doi.org/10.17605/OSF.IO/5BRXD

Kim, J., Byungkyu, L., (2023, Nov 11). *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction* https://arxiv.org/abs/2305.09620




DATA: 68,846 individuals' responses to 3,110 questions collected for 33 repeated cross-sectional data between 1972 and 2021 for fine-tuning the LLMs. Retrieved text content of GSS survey questions from GSS data explorer

**Figure 2: An overview of our methodological framework.** In Panel A, we use survey weights when aggregating individual-level prediction into population-level estimates to account for potential sampling bias. In Panel B, individual belief and period embeddings are initially randomly assigned but optimized during the fine-tuning process using dense and cross layers. Semantic embedding, initially estimated by pre-trained LLMs (e.g., Alpaca-7b), is also optimized during the fine-tuning stage.

Imputation is promising – what is needed is an evaluation comparing different imputation techniques.

Missing data always easier to handle when the mechanism for missingness is random/known.

(Potential) biases due to training data and instruction prompts should be kept in mind.

# Adaptive Matrix Sampling for the Consumer Expenditure Quarterly Interview Survey

Jeffrey M. Gonzalez
Gonzalez.Jeffrey@bls.gov*

John L. Eltinge†

**Abstract**

The Consumer Expenditure Quarterly Interview Survey is an ongoing panel survey of U.S. households in which sample units typically receive the same survey protocol during each interview. Because of the high burden associated with the survey request, the U.S. Bureau of Labor Statistics is exploring alternative designs that, if implemented, would change many features of the data collection process. One such alternative is adaptive matrix sampling. In general, matrix sampling involves dividing a survey into subsets of questions and then based on some probabilistic mechanism administering each to subsamples of the main sample. To potentially compensate for the resulting loss of information, as not all questions are asked of all sample units, we propose an adaptive assignment of subsampling probabilities based on data from the first interview. We use historical data to explore potential efficiency gains incurred by the use of this form of adaptive matrix sampling, develop point estimators based on simple weighting adjustments for expenditures collected under this design, and evaluate their variance properties.

**Key Words:** Adaptive design; Burden reduction; Multiple imputation; Sample survey; Two-phase sampling; Variance estimation

## 1. Introduction

### 1.1 The Consumer Expenditure Quarterly Interview Survey

The Consumer Expenditure Quarterly Interview Survey (CEQ) is an ongoing rotating panel survey of U.S. households in which, for each wave, all sample units are generally administered the same survey questionnaire. Each respondent is asked questions on a common set of expenditures. These expenditures are those that can be expected to be recalled for a period of three months or longer and tend to include relatively large purchases, such as for property and automobiles, and regularly occurring purchases, such as utility bills or insurance premiums. The data collected provide the basis for revising the weights and associated pricing samples for the Consumer Price Index (CPI), one of the nation's leading economic indicators, as well as a complete picture of a household's spending pattern (BLS *Handbook of Methods*, 2007).

The CEQ was designed as a personal visit interview and takes 50 to 60 minutes to complete depending on the interview. The preferred mode of data collection, at least from managerial and data quality perspectives, is personal visit; however, a substantial proportion of interviews are currently being conducted over the telephone. Safir *et. al.* (2008) point out that the percentage of cases completed by telephone has fluctuated over the years, but most recently has stabilized at about 35 percent. The increased practice of conducting CEQ interviews over the telephone has likely been made to mitigate unit nonresponse, but even so the response rate has been gradually declining over recent years. For example, response for the survey was about 80 percent in 2000, but by 2007, the annual response rate dropped to about 74 percent (BLS *Handbook of Methods*, 2007).

https://www.bls.gov/cex/cesrvymethsgonzale1.pdf

QUESTIONNAIRE DESIGN



SYNTHETIC DATA

INTERVIEWER

# A(I)utomatization in Classification

## occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys

**Jan Simson** [1¶], **Olga Kononykhina**[1], **and Malte Schierholz** [1]

**1** Department of Statistics, Ludwig-Maximilians-Universität München, Germany ¶ Corresponding author

**DOI:** 10.21105/joss.05505

**Software**

- Review
- Repository
- Archive

**Editor:** Chris Vernon
**Reviewers:**

- @welch16
- @danielruss

**Submitted:** 30 March 2023
**Published:** 24 August 2023



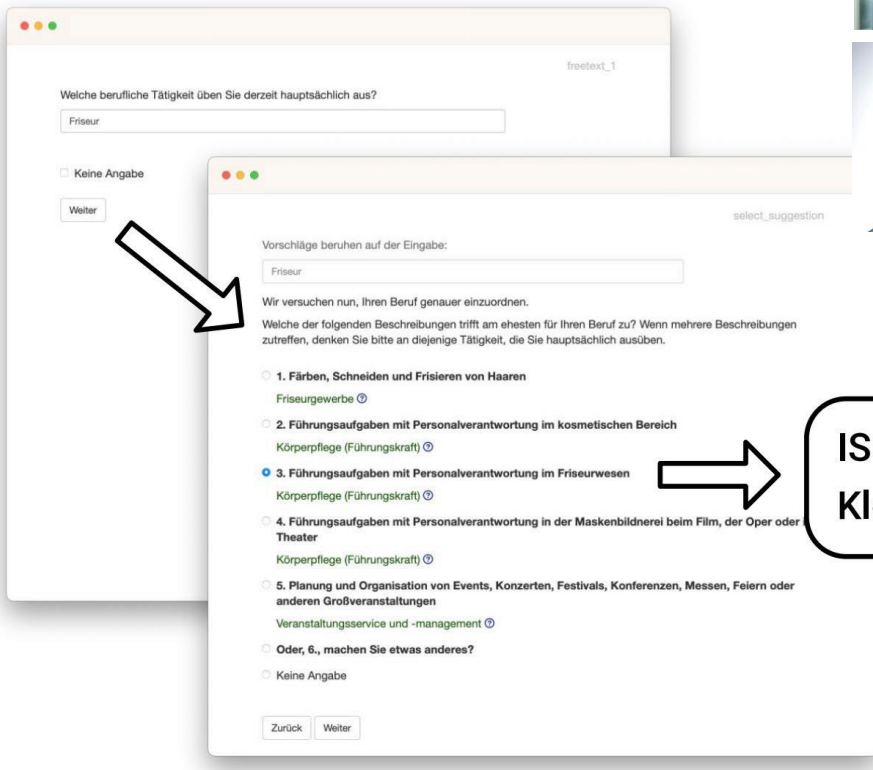
### Summary

People earn a living a multitude of ways which is why the occupations they pursue are almost as diverse as people themselves. This makes quantitative analyses of free-text occupational responses from surveys hard to impossible, especially since people may refer to the same occupations with different terms. To address this problem, a variety of different classifications have been developed, such as the International Standard Classification of Occupations 2008 (ISCO) (ILO, 2012) and the German Klassifikation der Berufe 2010 (KldB) (Bundesagentur für Arbeit, 2011), narrowing down the amount of occupation categories into more manageable numbers in the mid hundreds to low thousands and introducing a hierarchical ordering of categories. This leads to a different problem, however: Coding occupations into these standardized categories is usually expensive, time-intensive and plagued by issues of reliability.

Here we present a new instrument that implements a faster, more convenient and interactive occupation coding workflow where respondents are included in the coding process. Based on the respondent's answer, a novel machine learning algorithm generates a list of suggested occupational categories from the Auxiliary Classification of Occupations (Schierholz, 2018), from which one is chosen by the respondent (see Figure 1). Issues of ambiguity within occupational categories are addressed through clarifying follow-up questions. We provide a comprehensive toolbox including anonymized German training data and pre-trained models without raising privacy issues, something not possible yet with other algorithms due to the difficulties of anonymizing free-text data.

### Statement of Need

Assigning occupations to standardized codes is a critical task frequently encountered in research, public administration and beyond: They are used in government censuses (e.g. USA, UK, Germany) and administrative data to better understand economic activity, in epidemiology to estimate exposure to health hazards, and in sociology to obtain a person's socio-economic



# TOPCAT

Ma, B.; Haensch, AC; Resnick P.; Kreuter, F. (2024): Topic-Oriented Protocol for Content Analysis of Text – A Preliminary Study

Machine: analysis large quantities of data at scale
Human analysts: interpretation informed by their expertise and their knowledge of the questions the analysis is intended to help answer.

Human process is guided by the automatically proposed categories, it encourages inter-analyst reliability in the identification of categories.

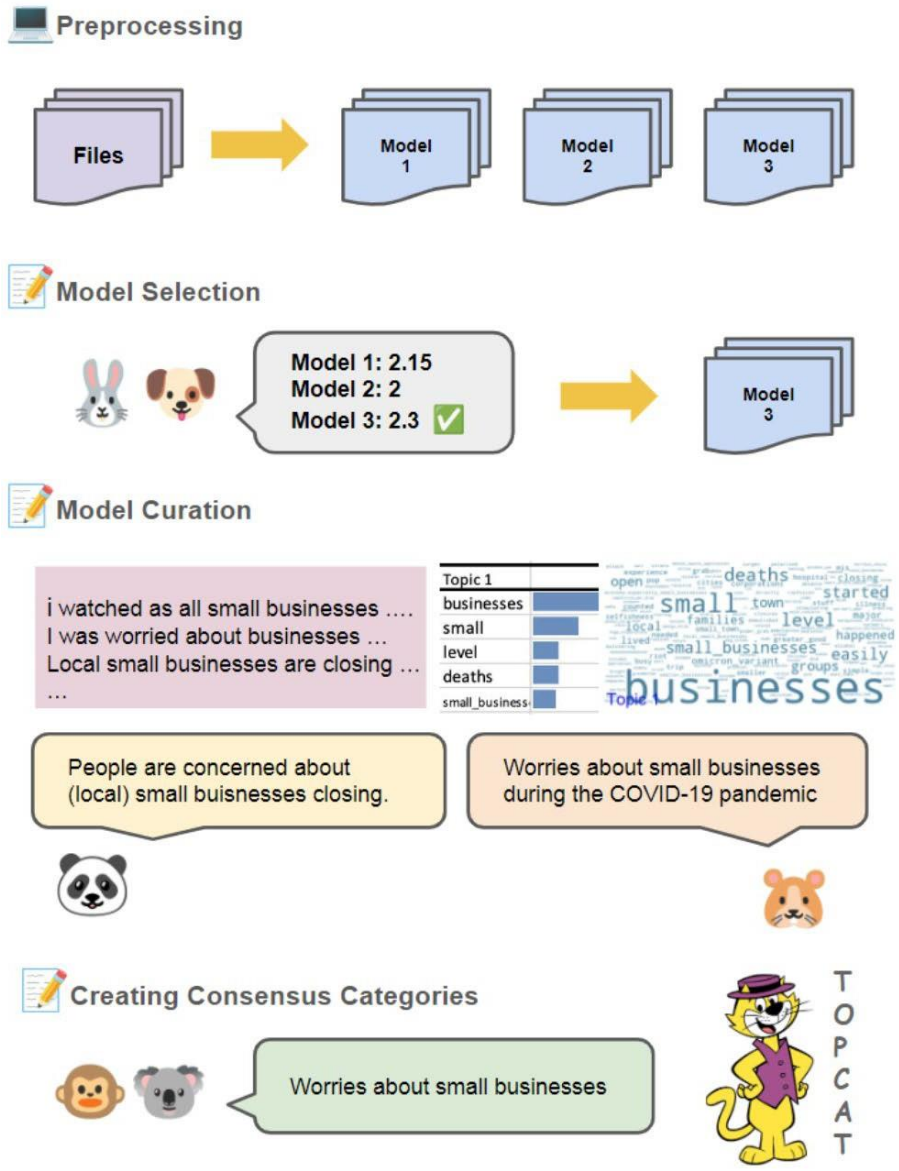



Figure 1: TOPCAT Workflow


Figure 1: Illustration of the concurrent interview settings (human- vs. AI-conducted) and the various metrics (👤, 👁, 📄 and 🔍) applied to assess interview quality.

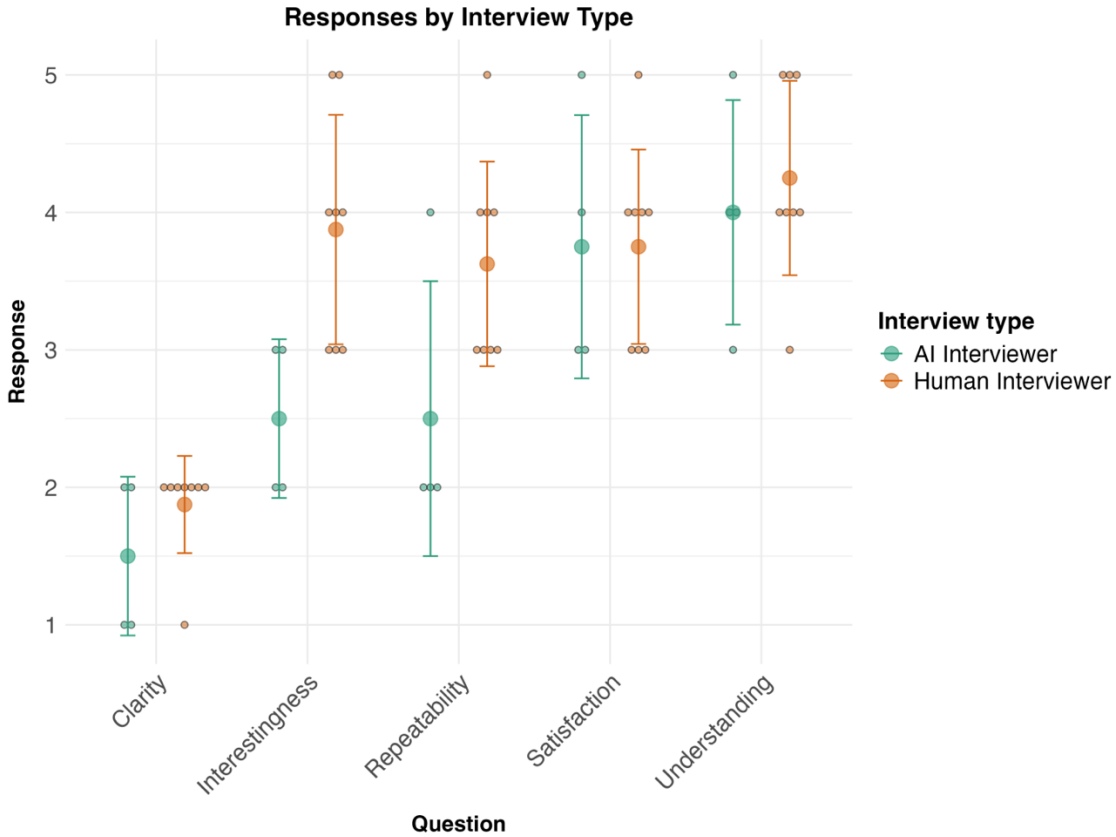
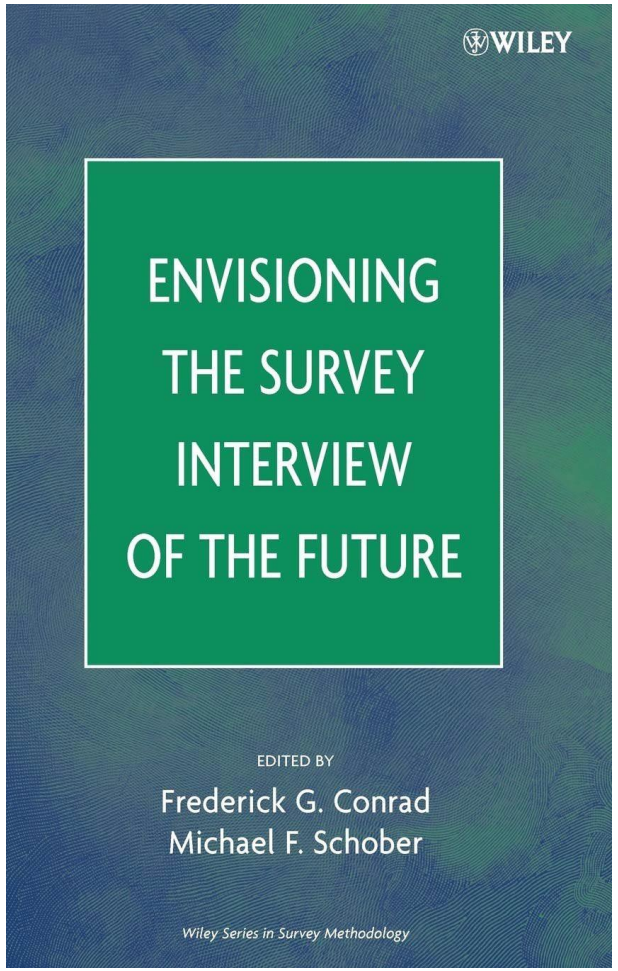



Figure 4: Participants' evaluation of interview 📄.

**AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers.**
Wuttke, Assenmacher, Klamm, Lang, Würschinger, Kreuter (2024)
https://www.arxiv.org/abs/2410.01824




- How and when should new communication technology be adopted in the interview process?
- What are the principles that extend beyond particular technologies?
- Why do respondents answer questions from a computer differently than questions from a human interviewer?
- How can systems adapt to respondents' thinking and feeling?
- What new ethical concerns about privacy and confidentiality are raised from using new communication technologies?

QUESTIONNAIRE DESIGN



SYNTHETIC DATA

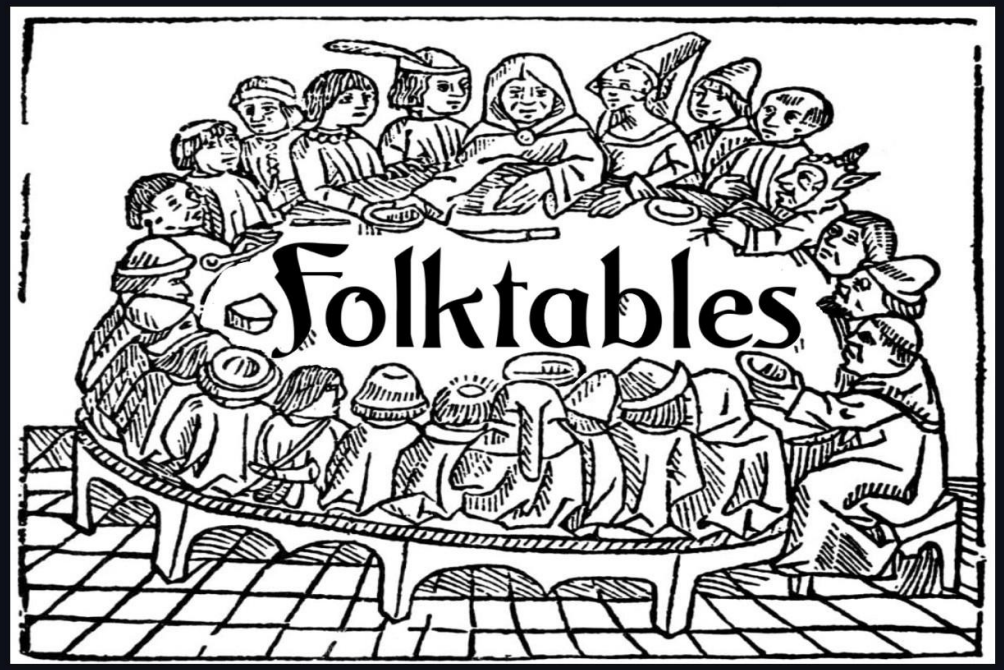


INTERVIEWER

# Leveraging Survey Research for AI

README.md

License MIT Downloads 18k pypi v0.0.12




**Folktables** is a Python package that provides access to datasets derived from the US Census, facilitating the benchmarking of machine learning algorithms. The package includes a suite of pre-defined prediction tasks in domains including income, employment, health, transportation, and housing, and also includes tools for creating new prediction tasks of interest in the US Census data ecosystem. The package additionally enables systematic studies of the effect of distribution shift, as each prediction task can be instantiated on datasets spanning multiple years and all states within the US.

Why the name? Folktables is a neologism describing tabular data about individuals. It emphasizes that data has the power to create and shape narratives about populations and challenges us to think carefully about the data we collect and use.

# GPT training pipeline

| stage | pretraining | instruction finetuning |
| --- | --- | --- |
| data | huge amounts of high-quality text from the internet | few high quality examples and/or comparison data |
| compute | millions of GPU hours | thousands of GPU hours |
| outcome | base model: GPT, LLaMa | ChatGPT, ChatLLaMa |

Critical for performance

Critical for performance and for building harmless & helpful assistants

# DCLM-Baseline pipeline

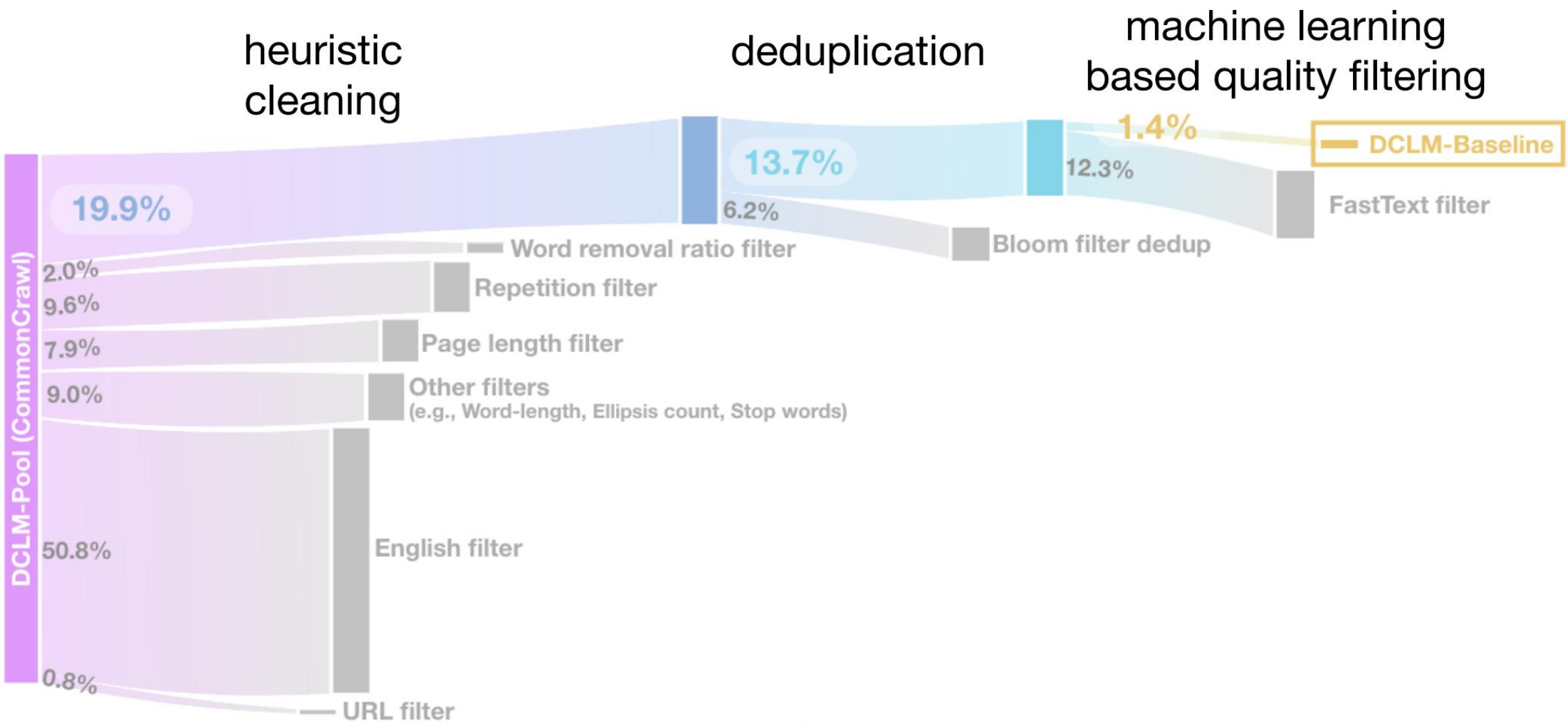


## DataComp-LM: In search of the next generation of training sets for language models

**Jeffrey Li*[1,2] Alex Fang*[1,2] Georgios Smyrnis*[4] Maor Ivgi*[5] Matt Jordan[4] Samir Gadre[3,6] Hritik Bansal[8] Etash Guha[1,15] Sedrick Keh[3] Kushal Arora[3] Saurabh Garg[13] Rui Xin[1] Niklas Muennighoff[22] Reinhard Heckel[12] Jean Mercat[3] Mayee Chen[7] Suchin Gururangan[1] Mitchell Wortsman[1] Alon Albalak[19,20] Yonatan Bitton[14] Marianna Nezhurina[9,10] Amro Abbas[23] Cheng-Yu Hsieh[1] Dhruba Ghosh[1] Josh Gardner[1] Maciej Kilian[17] Hanlin Zhang[18] Rulin Shao[1] Sarah Pratt[1] Sunny Sanyal[4] Gabriel Ilharco[1] Giannis Daras[4] Kalyani Marathe[1] Aaron Gokaslan[16] Jieyu Zhang[1] Khyathi Chandu[11] Thao Nguyen[1] Igor Vasiljevic[3] Sham Kakade[18] Shuran Song[6,7] Sujay Sanghavi[4] Fartash Faghri[2] Sewoong Oh[1] Luke Zettlemoyer[1] Kyle Lo[11] Alaaeldin El-Nouby[2] Hadi Pouransari[2] Alexander Toshev[2] Stephanie Wang[1] Dirk Groeneveld[11] Luca Soldaini[11] Pang Wei Koh[1] Jenia Jitsev[9,10] Thomas Kollar[3] Alexandros G. Dimakis[4,21] Yair Carmon[5] Achal Dave[†3] Ludwig Schmidt[†1,7] Vaishaal Shankar[†2]**

[1]University of Washington, [2]Apple, [3]Toyota Research Institute, [4]UT Austin, [5]Tel Aviv University, [6]Columbia University, [7]Stanford, [8]UCLA, [9]JSC, [10]LAION, [11]AI2, [12]TUM, [13]CMU, [14]Hebrew University, [15]SambaNova, [16]Cornell, [17]USC, [18]Harvard, [19]UCSB, [20]SynthLabs, [21]Bespokelabs.AI, [22]Contextual AI, [23]DatologyAI

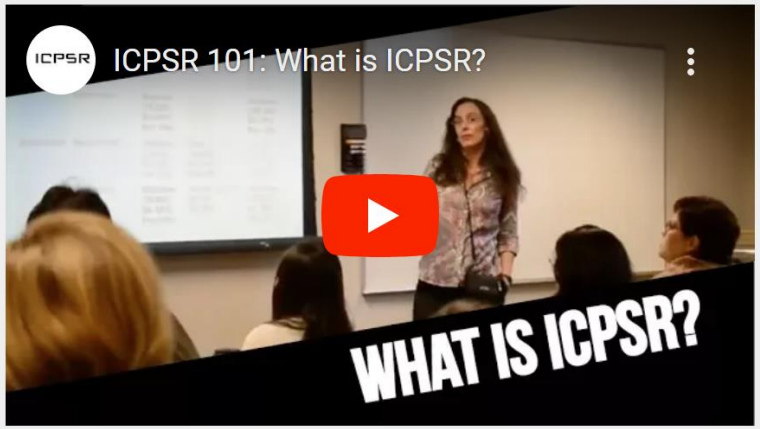
# Construction of DCLM-BASELINE



# About ICPSR

## Mission Statement

*ICPSR advances and expands social and behavioral research, acting as a global leader in data stewardship and providing rich data resources and responsive educational opportunities for present and future generations.*




ICPSR is an international consortium of more than 810 academic institutions and research organizations. ICPSR (Inter-university Consortium for Political and Social Research) provides leadership and training in data access, curation, and methods of analysis for the social science research community.

ICPSR maintains a **data archive** of more than 350,000 files of research in the social and behavioral sciences. It hosts 23 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields.

ICPSR collaborates with a number of funders, including U.S. statistical agencies and foundations, to create thematic data collections

## More information about ICPSR

- ICPSR receives grants from a number of government agencies and private foundations.
- A list of staff is available.
- The Consortium was established in 1962. Read about our history.
- ICPSR is governed by the ICPSR Council, a 12-person body elected by the members of ICPSR.
- ICPSR's governing documents include a constitution, bylaws, and a memorandum of agreement with the University of Michigan.
- ICPSR has annual reports dating back to 1962.

UK • DATA ARCHIVE

The UK's largest digital collection of social sciences and population research data

University of Essex | 60

WHERE CHANGE HAPPENS 1964–2024



Home to the UK's largest collection of social, economic and population data for over 50 years, we provide researchers with training, support and data access as lead partner of the UK Data Service.

## Managing data

Learn from our trusted international best practice

## Data Catalogue

Browse the largest collection of digital research data in the social sciences through the UK Data Service

## Resources

Free webinars, on-demand tutorials and resources to help improve your data skills

Services →

## → Planning studies and collecting data

- → Survey Methods Consulting
- → Questionnaire Development
- → Sampling
- → GESIS Panel
- → Tools for Collecting Digital Behavioral Data

## → Finding and accessing data

- → ALLBUS
- → Eurobarometer
- → EVS
- → GLES
- → ISSP
- → PIAAC
- → Election studies
- → International Survey Programs
- → GESIS Web Data

## → Processing and Analyzing Data

- → Weighting and Analysis of Complex Samples
- → Data harmonization
- → Service for Official Microdata
- → Analysis of Sensitive Data
- → Analyzing Digital Behavioral Data

# PARTNERSHIP CONSIDERATIONS

## Data Sharing

Representatives from Meta stated that their approach to research data sharing has evolved over the last ten years. Product teams and cross-functional teams (legal, policy, academic partnerships, etc.) work together to enable data sharing. They communicated that there are four main stages for data sharing; 1. identifying researcher needs, 2. understanding how to ensure user privacy and data security, 3. building data sets, and 4. maintaining data sets. By starting with identifying researcher needs, they say they try to efficiently meet those needs while building something of value for the research community. Additionally, their work centers on user privacy while attempting to identify interesting data sets or increase data utility.

The team remarked on misconceptions that sharing data is easy, explaining that building data sets for sharing is a fairly complex process. They added that it isn't as simple as just running an SQL query to produce a data set ready to be shared. Oftentimes they have to combine data sets in specific ways to pass internal quality assurance requirements, and each process usually involves new work. If the team determines that the data they created is of sufficient quality and accuracy

## Data Sharing Agreement

Meta representatives described the use of multiple forms of data sharing agreements (DSAs) depending on the type of partnership being considered. They work with researchers' institutions to ensure DSAs meet the needs of everyone involved. Meta leveraged **Social Science One** in its effort to negotiate a **standard DSA** for researchers to request Facebook data for certain research questions. The data sharing team expressed support for the European Digital Media Observatory's (**EDMO**) working group's approach to data sharing agreements. Additionally, the Inter-university Consortium for Political and Social Research (**ICPSR**) agreed to host data from Facebook and Instagram related to the **US 2020 Election** and has its own DSA to which researchers requesting access to data must agree. Their DSAs also address scientific oversight, an area where 3rd parties can be useful. If researchers want to use sensitive data in a publication, Meta can stipulate that it can review the data prior to publication to ensure user privacy isn't compromised.

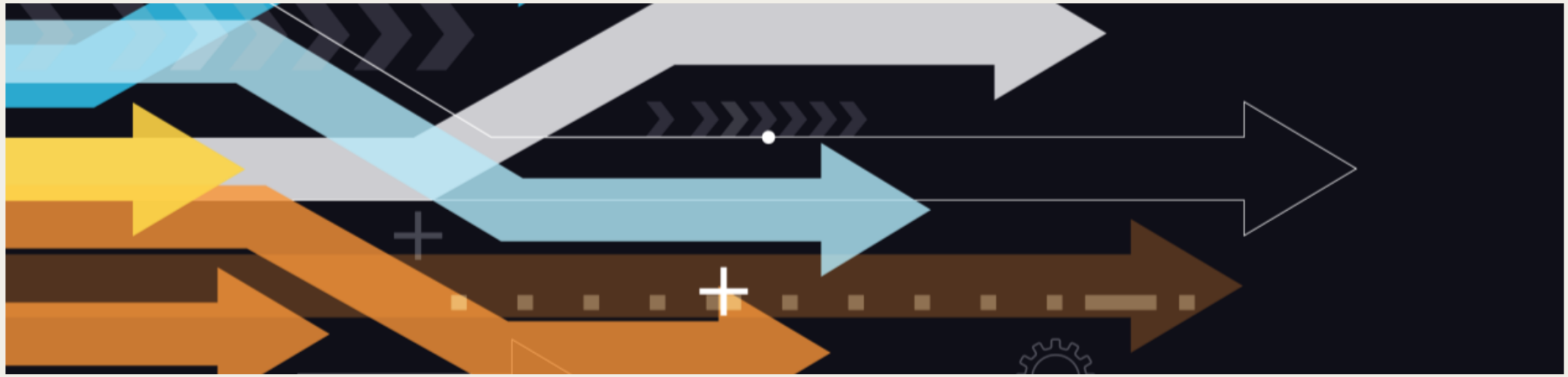
## Data Sharing Frequency

Representatives communicated that they regularly engage in data sharing with researchers, but the frequency depends on the project. For example, their **Meta ads library**, a dataset of

https://fpf.org/wp-content/uploads/2023/08/FPF-Data-Sharing-Case-Study-Meta-R2.pdf

# A Workshop on Future Directions for Social and Behavioral Science Methodologies in the Next Decade







The workshop gathers a broad group of experts to explore methodological and analytical innovations in the social and behavioral sciences, focusing discussions on future needs and methodological frontiers that are expected to benefit more than one discipline. Spanning the full data lifecycle, the workshop considers developments in artificial intelligence, machine learning, spatial analysis, causal modeling, survey methods, and the utilization of various data sources for social, behavioral, and economic research. The workshop is

AGENDA

DATE(S)

Sep 25 - 26, 2024

NATIONAL ACADEMIES Sciences Engineering Medicine

About Us Events Our Work Publications

# A Workshop on Future Directions for Social and Behavioral Science Methodologies in the Next Decade

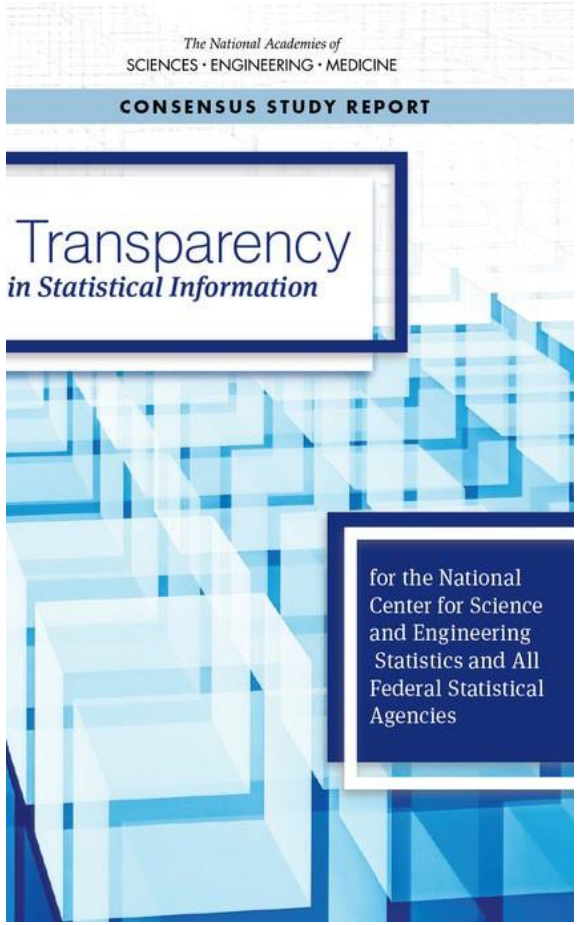

What guarantees do you think we need in order to know we have a result (worth publishing)? (Threat: Junk Science)

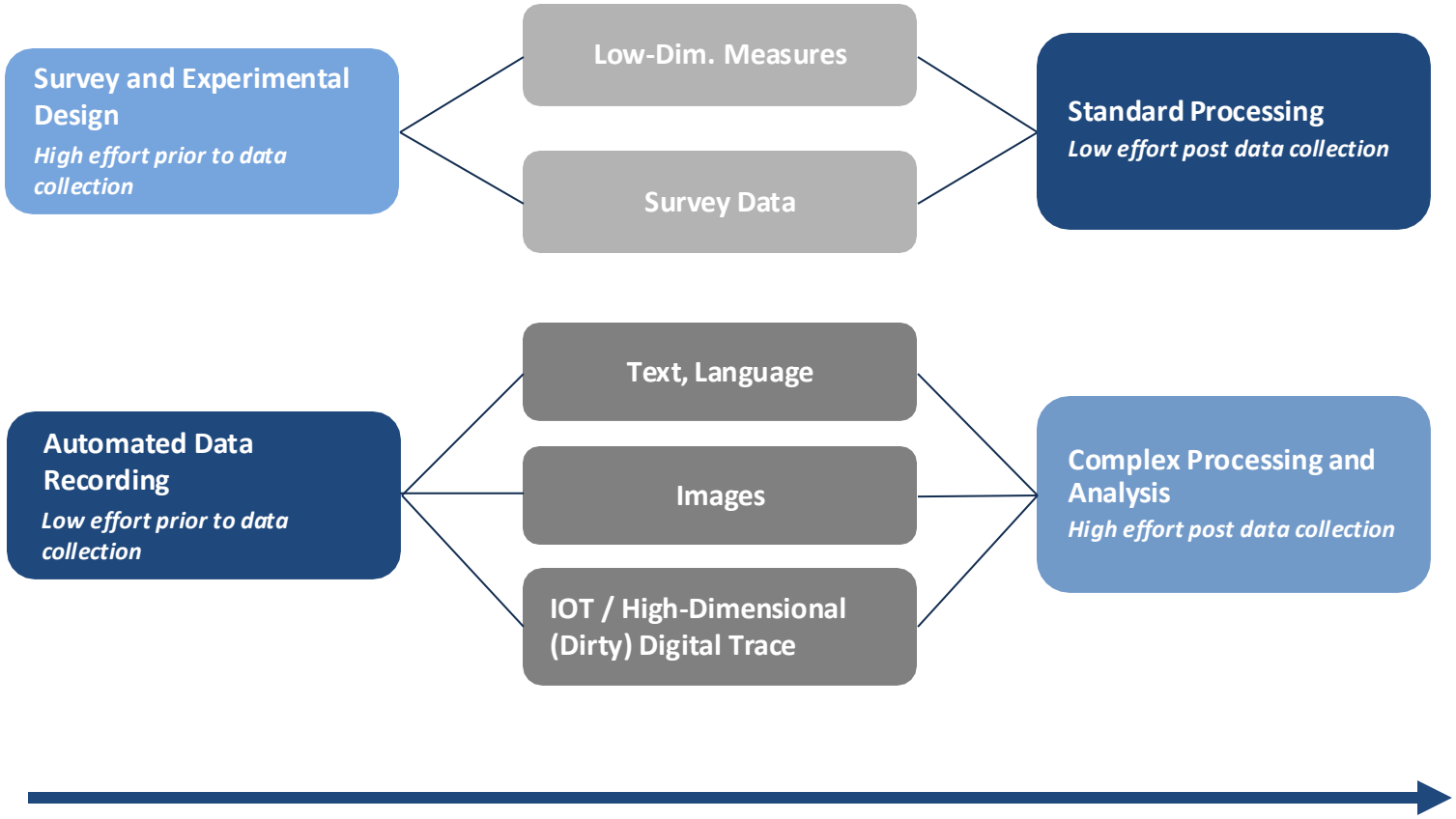
What method description is the minimum you would want to see when reviewing papers? (Threat: Bias, Interpretability, Robustness)

How do we bring proposal reviewers up to speed so they can reasonably evaluate the suggested research proposals?

How can TCS contribute to improving the alignment, fairness, and reliability of LLMs in social science research?

# Transparency is a Challenge



# Transparency Initiative


JUMP TO: Select Section

## What is the Transparency Initiative?

AAPOR's Transparency Initiative is designed to promote methodological disclosure through a proactive, educational approach that assists survey organizations in developing simple and efficient means for routinely disclosing the research methods associated with their publicly-released studies.

The Transparency Initiative is an approach to the goal of an open science of survey research by acknowledging those organizations that pledge to practice transparency in their reporting of survey-based research findings. In doing so, AAPOR makes no judgment about the approach, quality or rigor of the methods being disclosed.

## Join the TI! You will be in great company!

- Why should my organization join?
- How does my organization join?
- Where can I get more information?


# AAPOR 80th Annual Conference

Call for Abstracts

Schedule at a Glance

## The AAPOR Annual Conference is the premier forum for the