Leveraging AI for Survey Research

Frauke Kreuter
JPSM and SoDa Center, University of Maryland & LMU Munich
fkreuter@umd.edu

Menti.com use code 3935 9791

How would you use AI in survey research?

- *Designing the survey instrument* -- LLMs could help phrase the questions and pinpoint any inconsistencies
- *Sampling* — LLMs could suggest appropriate samples and techniques for recruiting participants.
- *Data cleaning and management* — LLMs could, perhaps, detect inconsistent and uniform selections, …identifying gibberish and spelling mistakes in open-ended responses.
- *Data analysis* - .. ChatGPT's Code Interpreter plugin to automate data analysis
- *Reporting and dissemination* -- LLMs that can implement data science code could help facilitate this process.

Using ChatGPT to Write Survey Questions

Survey Research

ChatGPT has received a lot of attention since its release as it can solve so many problems for people in range of fields, from research to development.

What is ChatGPT?

ChatGPT is a language model developed by OpenAI, trained to generate human-like text based on a g prompt.

There's a lot you can use ChatGPT for, including:

Ask ChatGPT to recommend question types

If you're a researcher or survey administrator, you'll know that choosing a question type is equally as important as the questions themselves.

Well, ChatGPT will recommend a question type based on your question text.

Prompt: What question type should I use for this survey question: How satisfied are you with the overall performance of our new product?

ChatGPT output:

> The question "How satisfied are you with the overall performance of our new product?" is a question that is best suited for a Likert scale, which is a type of rating scale that allows respondents to indicate their level of agreement or satisfaction with a statement or question. It is a commonly used question type in customer satisfaction surveys. The scale usually ranges from Strongly disagree to Strongly Agree or similar. In this case, you can use a 5-point Likert scale (1- strongly disagree, 2-disagree, 3-neutral, 4-agree, 5- strongly agree) to measure the level of satisfaction of the customers.

Showing results for designing questionnaires with ***chatgpt***
Search instead for designing questionnaires with chaptgpt

Videos

How to design a survey using ChatGPT

YouTube · Mohamed Benhima
Apr 16, 2023

How to generate survey questions for research using ChatGPT

YouTube · Tasvir Mahmood
Mar 26, 2023

Using ChatGPT to design a research study: research question ...

YouTube · Ranywayz Random
Jan 28, 2023

10 key moments in this video

How to use ChatGPT to create survey items: 3 examples ...

YouTube · Ranywayz Random
Feb 11, 2023

3 key moments in this video

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.

10 survey questions to capture students' satisfaction

On a scale of 1-10, how satisfied were you with this class?

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)

How effective was the instructor in delivering the course material?
(1 = Not effective at all, 10 = Extremely effective)

How encouraged did you feel to participate and engage in class discussions?
(1 = Not encouraged at all, 10 = Highly encouraged)

Were the provided class materials (e.g., textbooks, handouts, online resources) helpful in understanding the course content?
(1 = Not helpful at all, 10 = Extremely helpful)

Here's a revised set of questions with explanations for each improvement based on insights from Schuman and Presser

On a scale of 1-10, how satisfied were you with this class?
Revised: How would you describe your overall satisfaction with this class?
Options: 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.

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.

- We have 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 survey items/response options output by the LLMs.

Experiment 4: Controlling the content and number of response options that are generated.

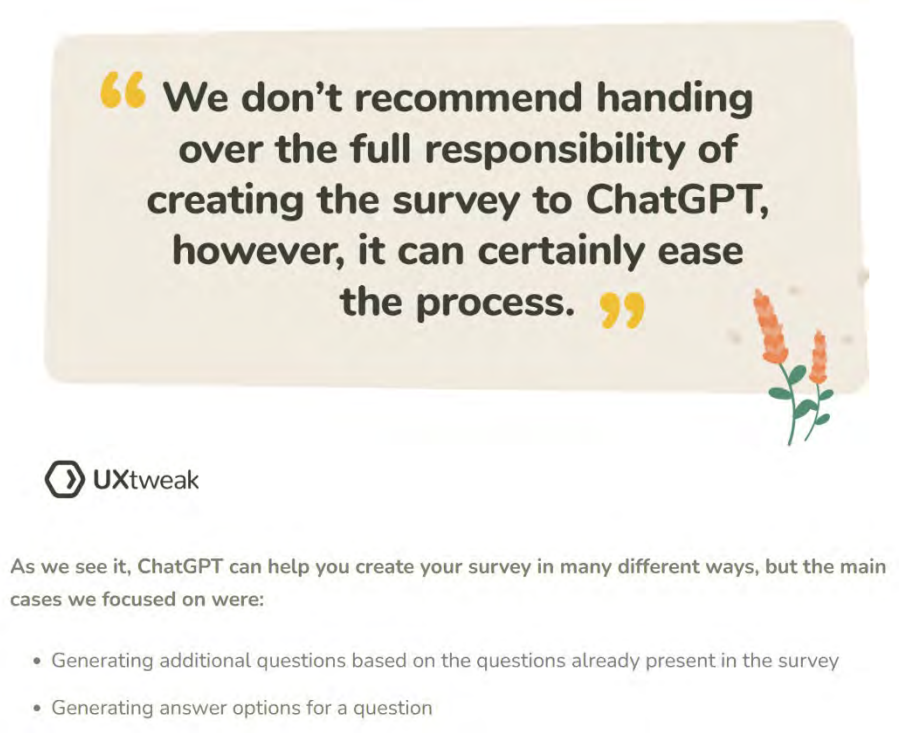
Some High Level Findings – Buskirk et al. 2024

- We have completed experiments 1 and 2 and are processing those results now. We are using a combination of Questionnaire Item Recommendations from Dillman, Smyth & Christian (2014) to guide our evaluation of outputs.
- High level learning from Experiment 1 has generated pretty clear signals:
  - Without using the word “survey” or “response options/answer choices” you get nearly 100% open-ended questions that appear more in third person and seem like college essay type questions.
  - If you use the phrase “answer choices” without “survey” when requesting a question from LLMs you seem to more than not get Trivia or Quiz type questions.
  - Including the word “survey” in the prompt generates mostly first-person oriented questions.
  - Adding “response options” to “survey” seems to generate first-person, closed-ended survey questions. However, many of these are imperative statements rather than direct questions.

More Results SOON!

- We are compiling more results now and will present them at the AAPOR conference on
  May 16, 2024 @1:45pm in Room 222.
- Hope to see you there!

- tdbuskirk@odu.edu



LLMS AND SYNTHETIC SAMPLES

A(I) Creating Data

Political Analysis

Article contents

Abstract

Footnotes

References

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

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]

Notes: We decided not to include "gewählt" (voted) as a suffix in the prompt, using the [MASK] instead of [INSERT] request, as it might bias the output against non-voters by reducing the likelihood of GPT completing the sentence with "nicht" (not) or "ungültig" (invalid) due to German semantics. We leave the further exploration of these effects to prompt engineering researchers.

	glous], sehr religios [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 [mod erately weakly
partyid	mit der mit d der Pa [Green mit der Partei [small with ar
east	0 West Ostdeu
immigration	erleich ichtern nor], ei
inequality	Maßna habe Regier [no op greifen

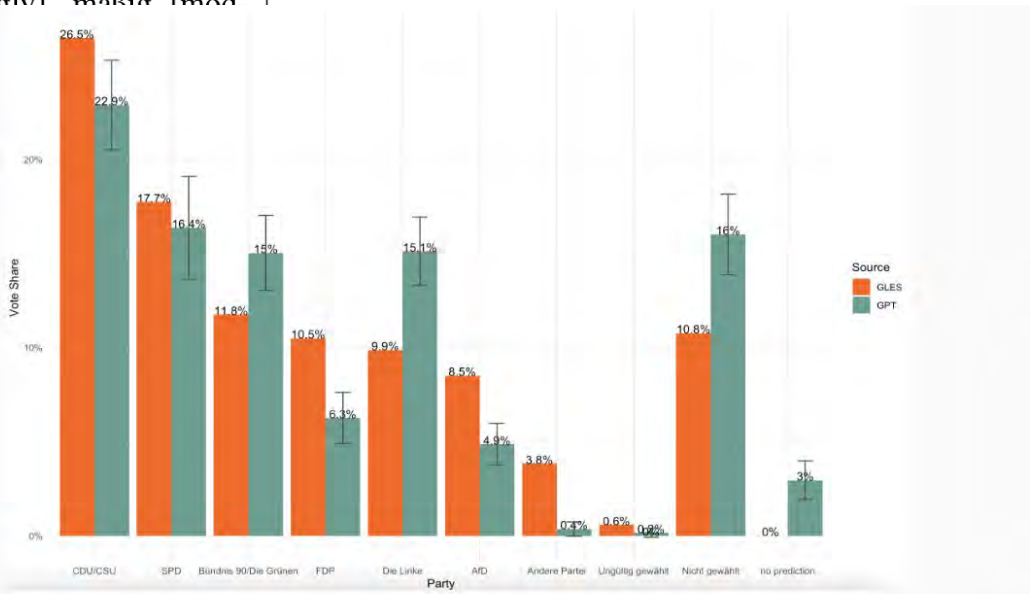


Figure 3: Replicating Argyle et al. for German data (GLES): Current project by Leah von der Heyde, Alexander Wenz and Carolina Haensch

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



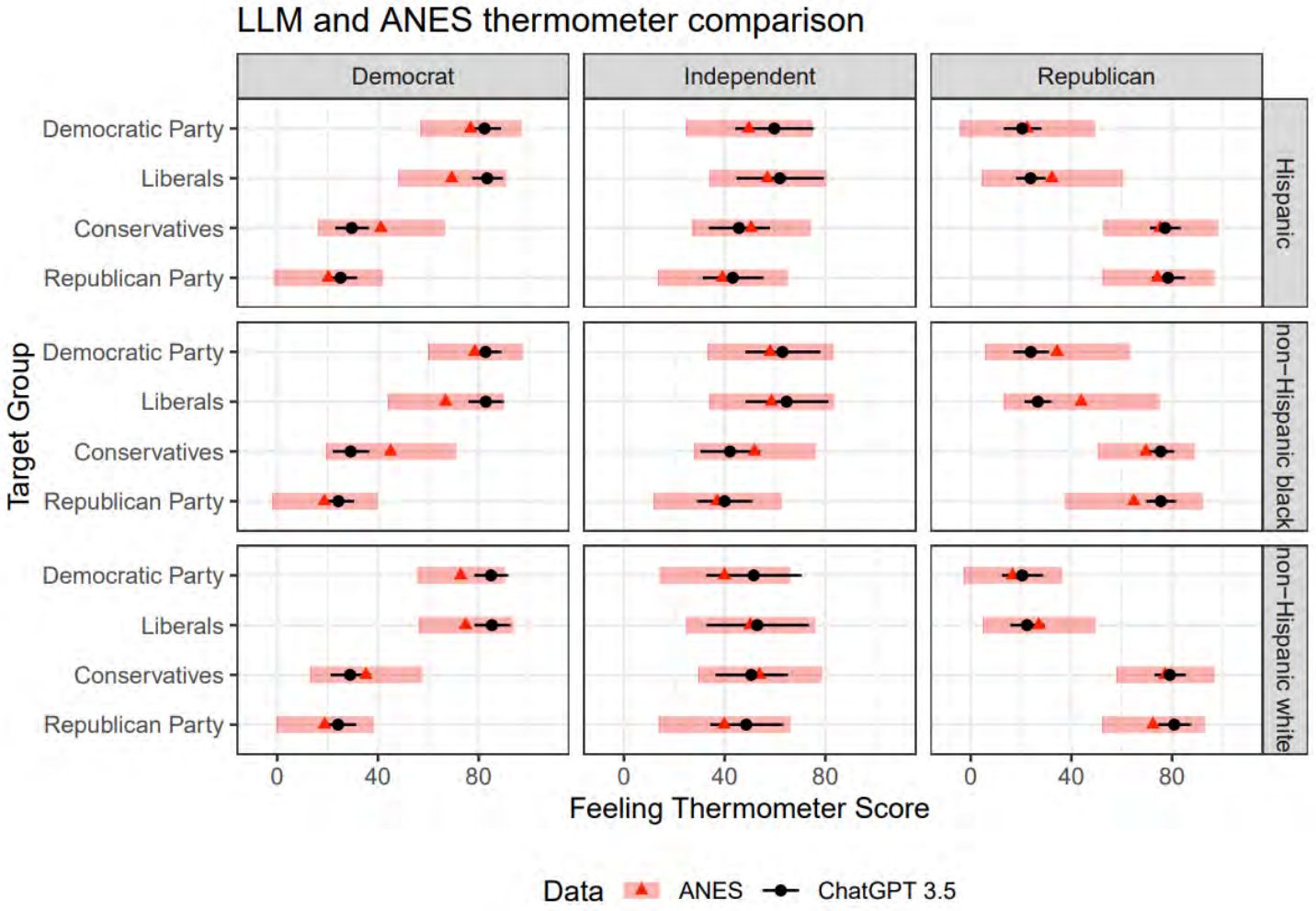


Figure 2: Average feeling thermometer results (x-axis) for different target groups (y-axes) by party ID of respondent (columns). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated by black circles and thin black bars. Sample sizes for each group-wise comparison are identical.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023, May 4). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. https://doi.org/10.31235/osf.io/5ecfa

LLMS FOR IMPUTATION

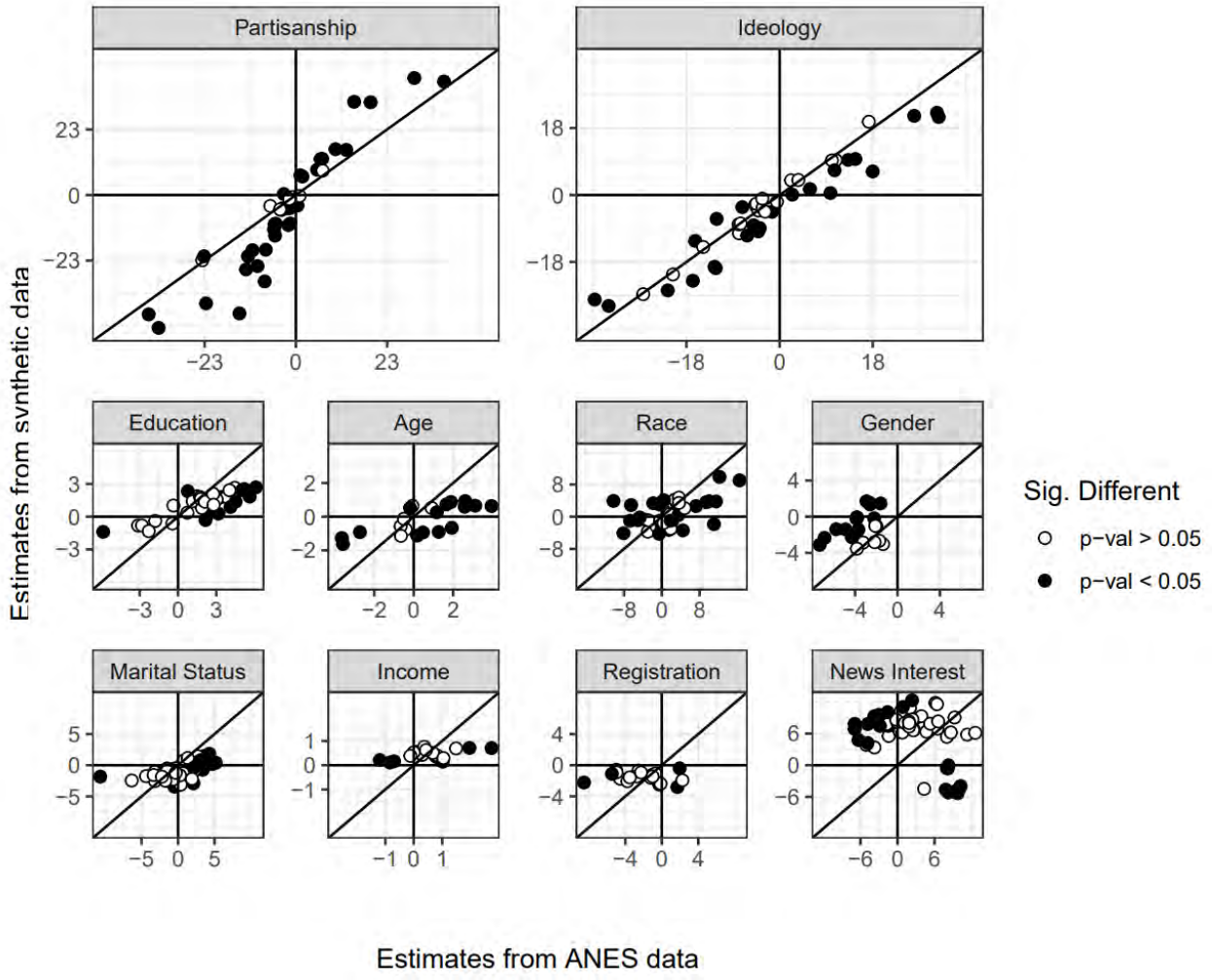


Figure 3: Each point describes the coefficient estimate capturing the partial correlation between a covariate and a feeling thermometer score toward one of the target groups, estimated in either 2016 or 2020. The x-axis position is the coefficient estimated in the ANES data, and the y-axis position is the same coefficient estimated in the synthetic data. Solid points indicate coefficients who are significantly different when estimated in either the ANES or synthetic data, while hollow points are coefficients that are not significantly different. Points in the northeast and southwest quadrants generate the same substantive interpretations, while those in the northwest and southeast quadrants produce differing interpretations. A synthetic dataset that is able to perfectly recover relationships estimated in the ANES data would have all points falling along the 45 degree line.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023, May 4). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. https://doi.org/10.31235/osf.io/5ecfa

LLMS AND MATRIX DESIGN

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

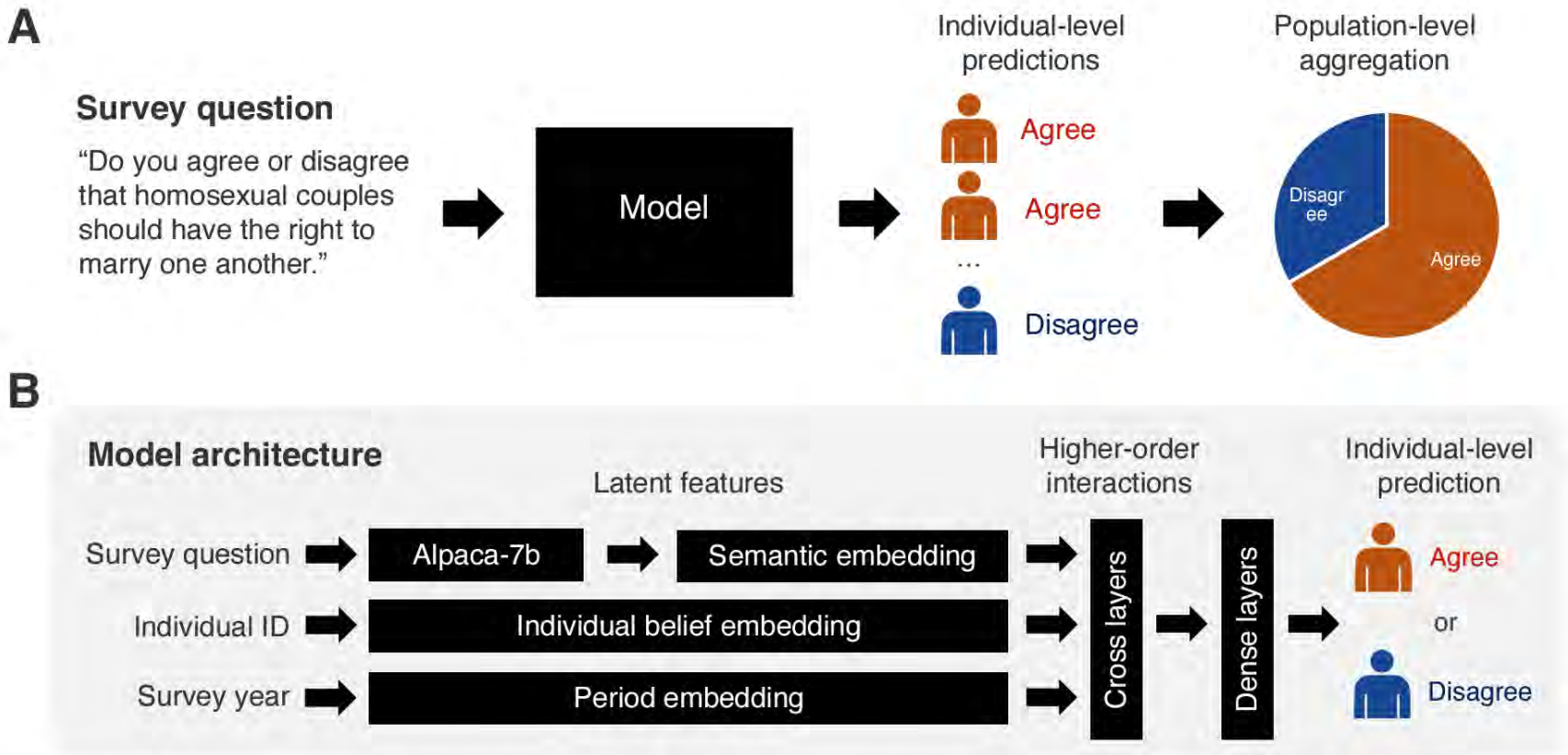


Figure 2: An overview of our methodological framework. In Panel A, we use survey weights when aggregating individual-level prediction into population-level estimates to account for potential sampling bias. In Panel B, individual belief and period embeddings are initially randomly assigned but optimized 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.

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

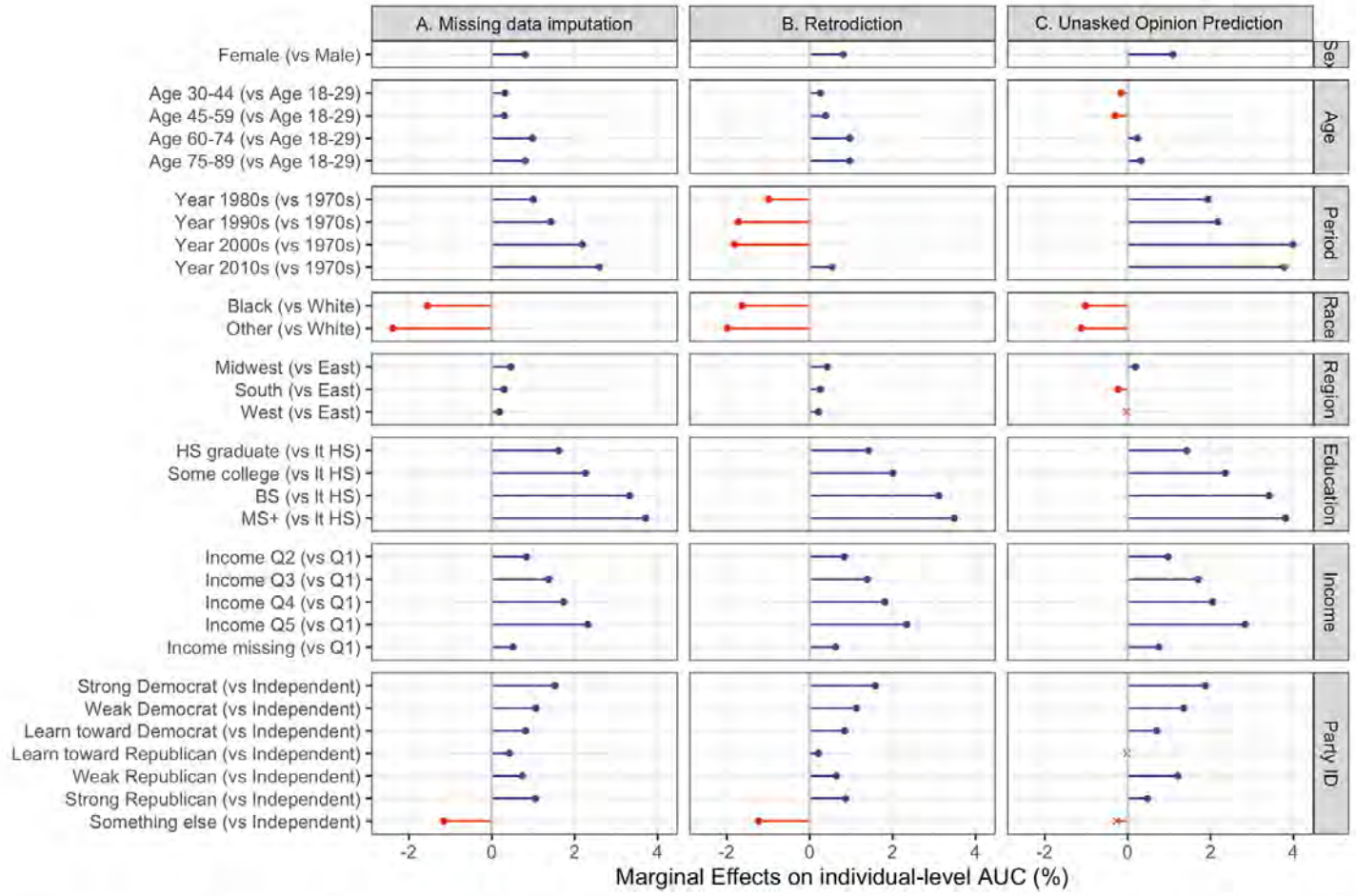


Figure 5: Coefficient plots from OLS regression models predicting individual-level AUC across three different types of missing response prediction. A higher AUC value indicates

For instance, rather than asking the same ten questions to a thousand participants, pollsters can disseminate twenty questions among the same thousand participants, each answering ten questions, and employ the model to infer individual responses to the remaining ten unasked questions. On the other hand, given our model's remarkable ability to mimic human responses, even including biases, researchers can use it to refine their survey questions by systematically examining characteristics of questions that cannot be accurately predicted (e.g., poor question wording).

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

ML AND LLMS FOR CODING TASKS

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

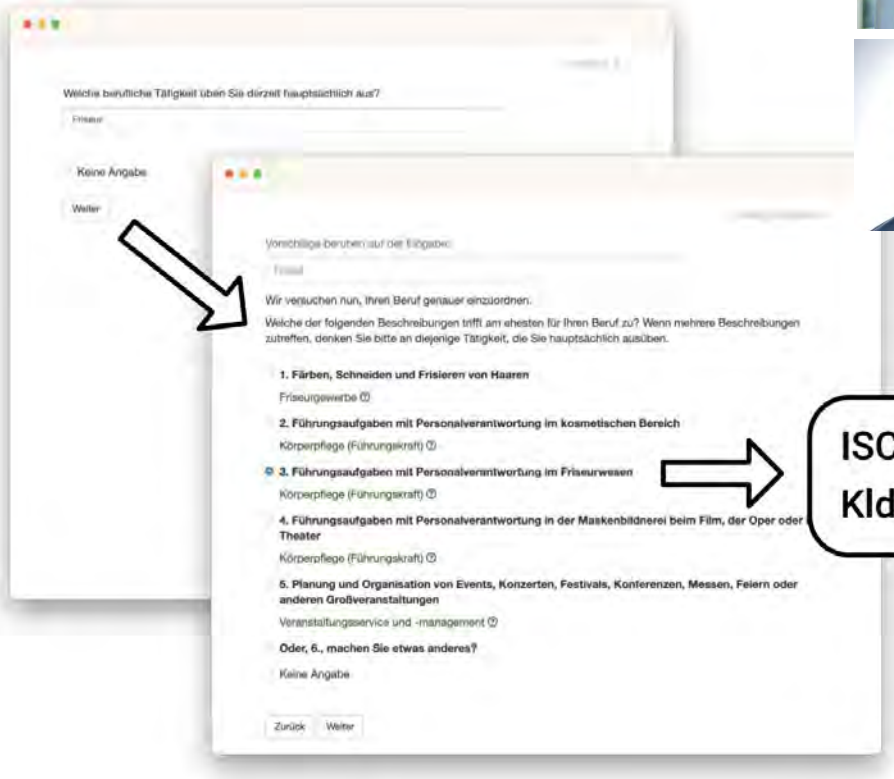
Summary

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.

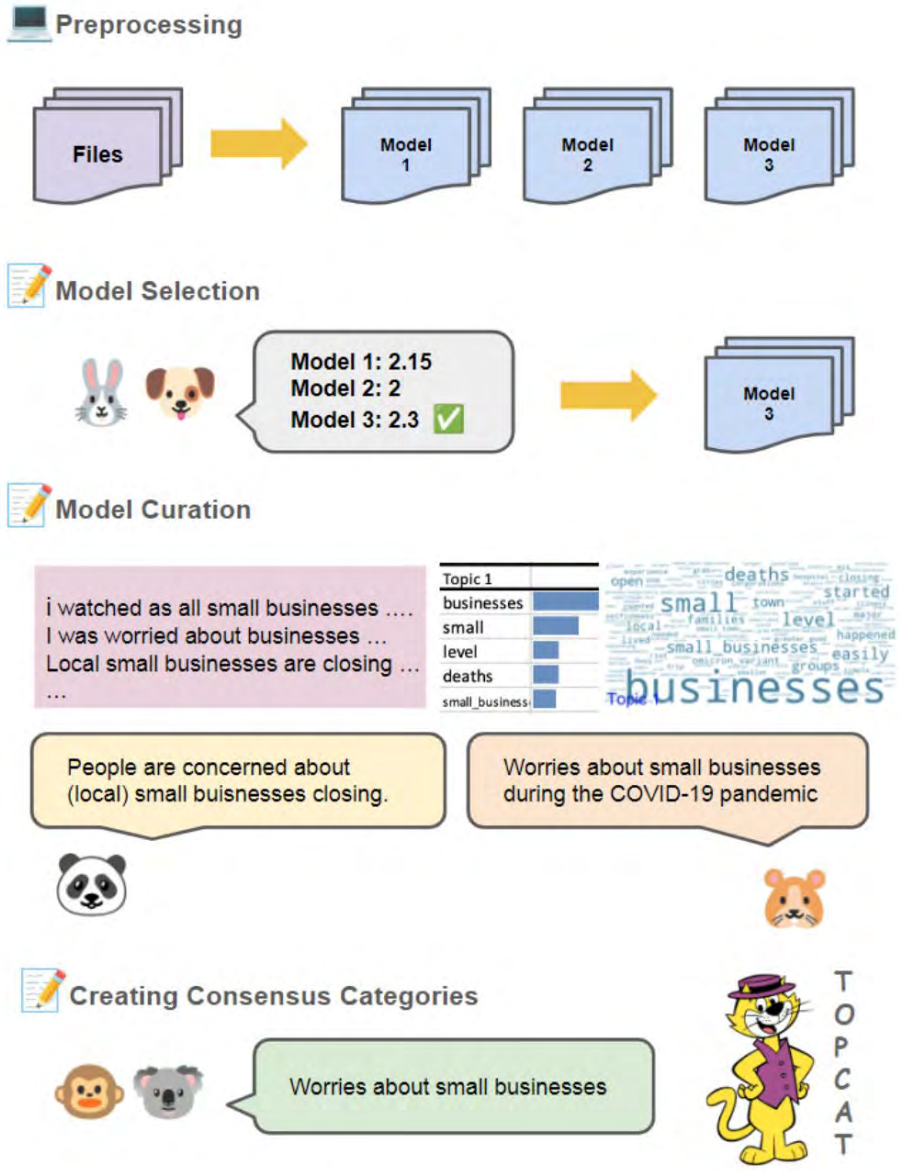
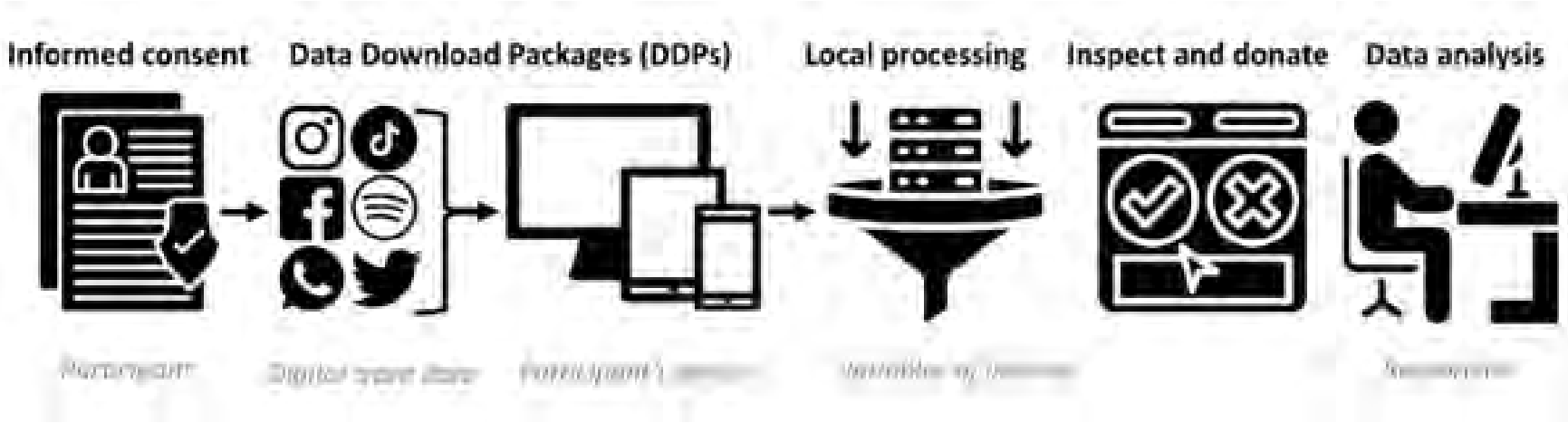


Figure 1: TOPCAT Workflow

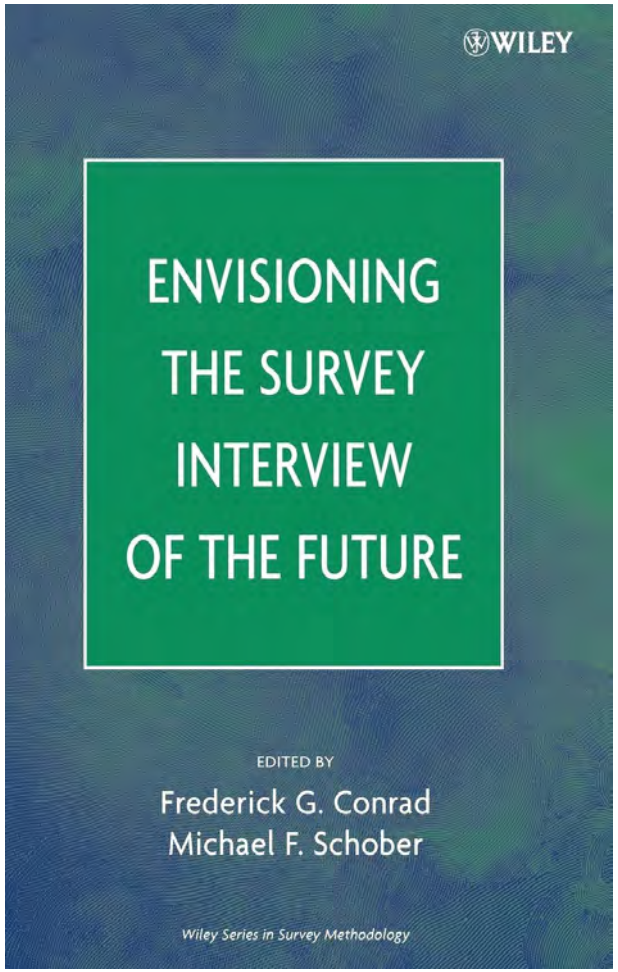
LLMS AS ANALYTIC ASSISTANT

Data Donation



https://datadonation.eu/

LLMS AS INTERVIEWER



- 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?

A(I)utomatization in Data Collection

Max Lang,
LMU Munich

Chainlit Chat Readme New Chat

You 00:45

6

InterviewGPT 00:45

I see that you have a high level of interest in politics, rating it as 6 out of 7. Could you elaborate and explain why you're so interested in politics? What aspects of politics engage you the most?

0:00 / 0:12

You 00:45

Social justice

InterviewGPT 00:46

It's insightful that social justice is a driving factor for your interest in politics. This implies you see politics as a means to achieve fairness and equality in society. Can you perhaps share your thoughts on what "politics" is? How would you define this term?

0:00 / 0:16

Type your message here...

Built with Chainlit

LEVERAGING SURVEY RESEARCH FOR AI

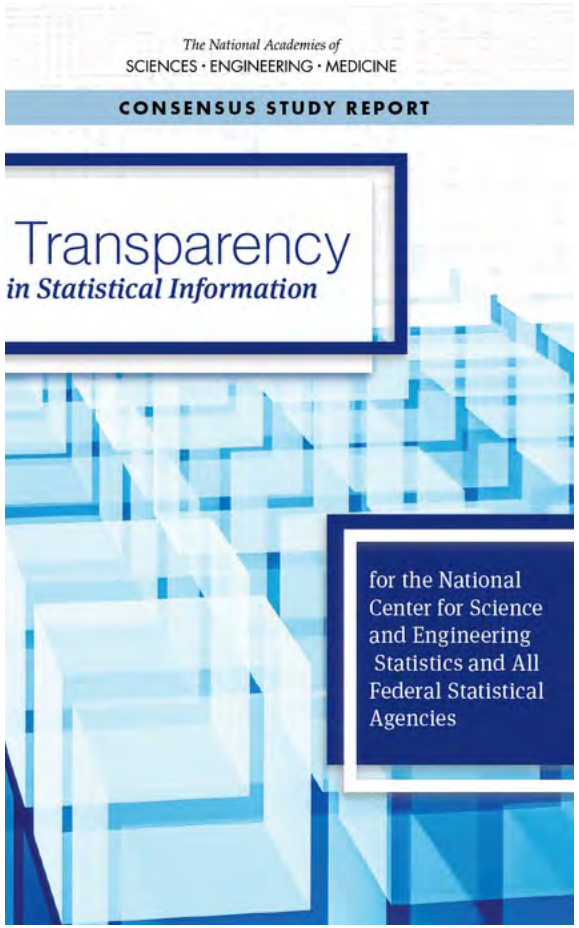
README.md

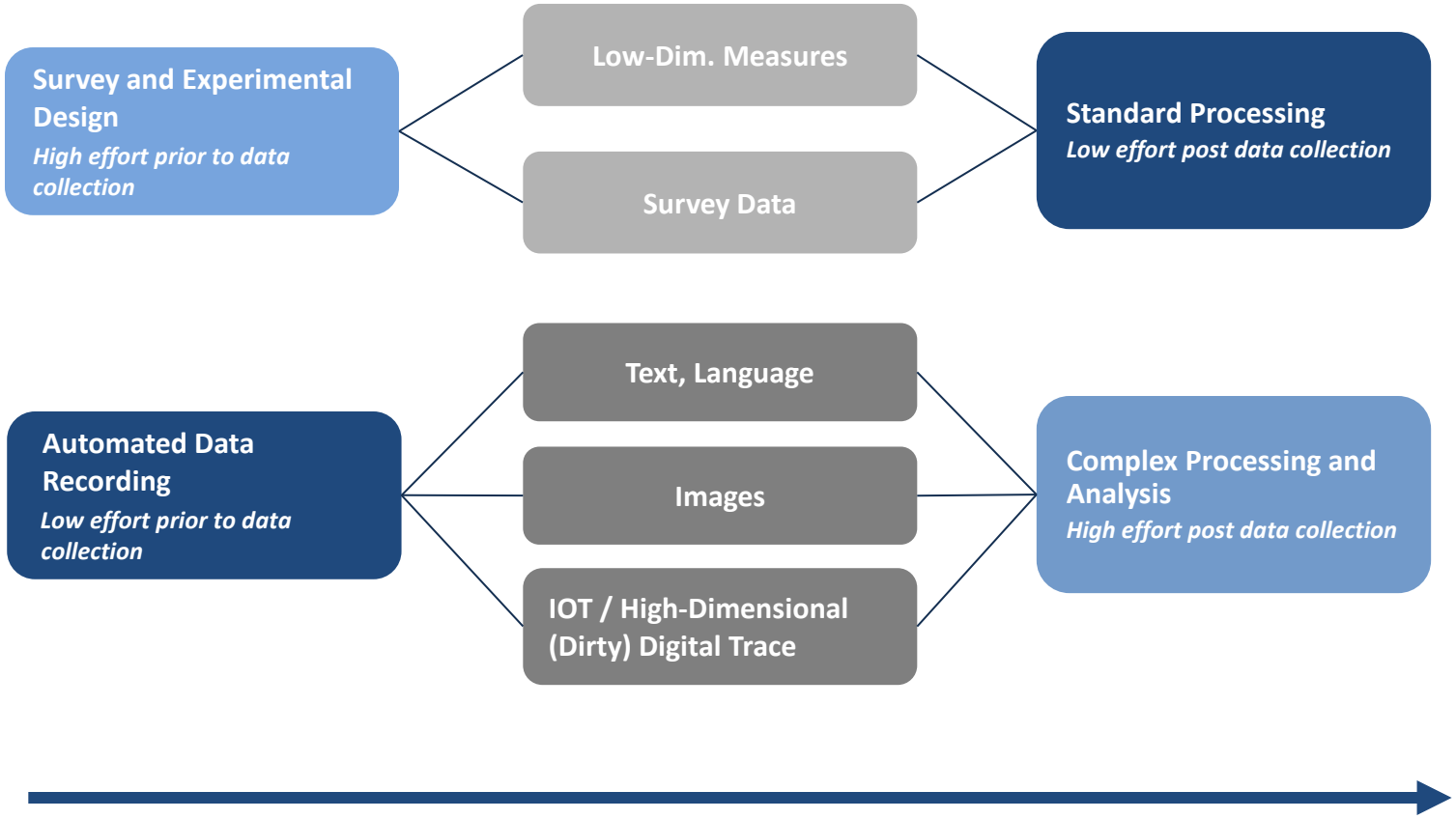
License MIT Downloads 18k pypi v0.0.12

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.

Transparency is a Challenge





BRINGING SURVEY METHODOLOGY TO MACHINE LEARNING

Stephanie Eckman Christoph Kern, Jacob Beck, Bolei Ma, Rob Chew, Frauke Kreuter



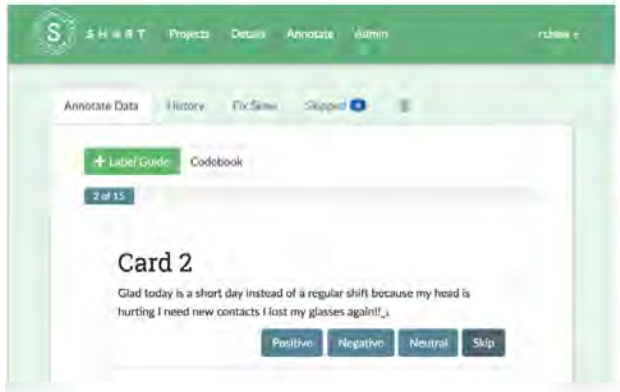


"The bias I am most nervous about is the bias of the human feedback raters"

Sam Altman
March 25 2023 "The Lex Fridman Podcast"

Annotation Sensitivity

Data Collection



Error Sources

- Nonresponse
- Order Effects
- Interviewer Effects

Beck et al (2022):

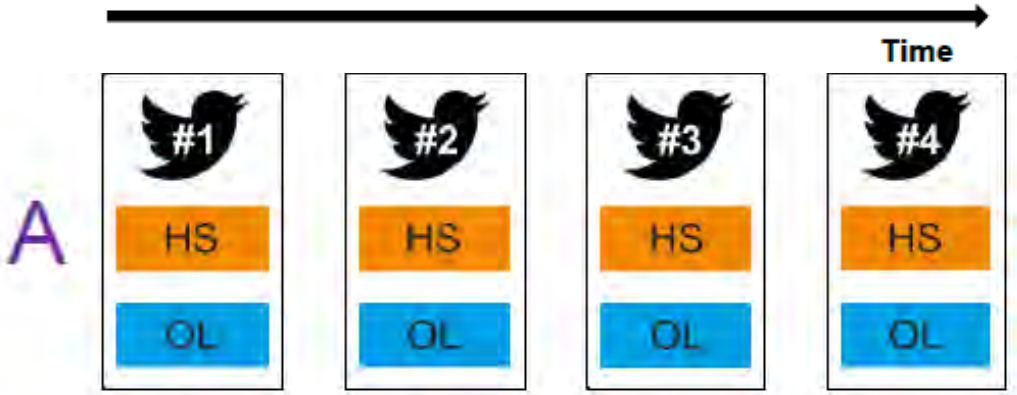
- Wording Effects
- Order Effects
- Annotator Effects

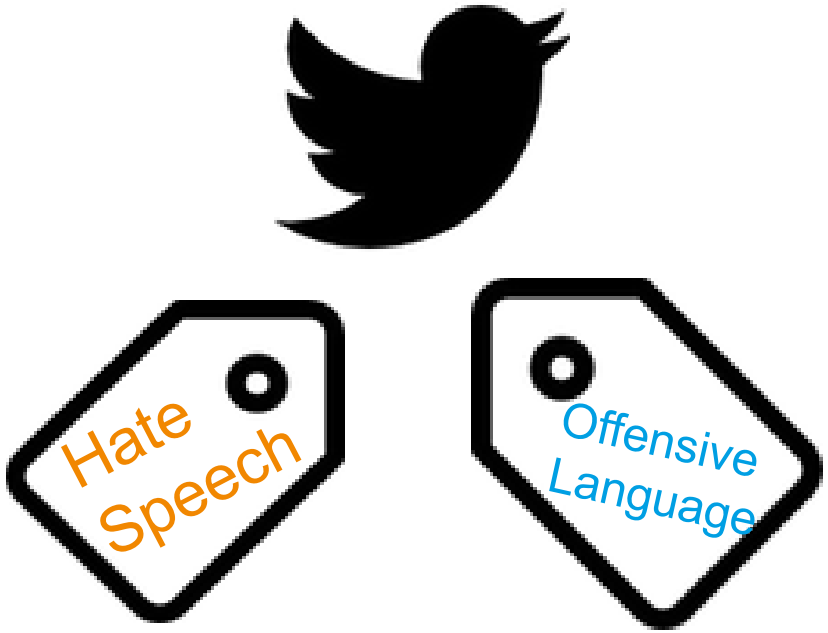
Impact

Bias

Prediction Error

Research design



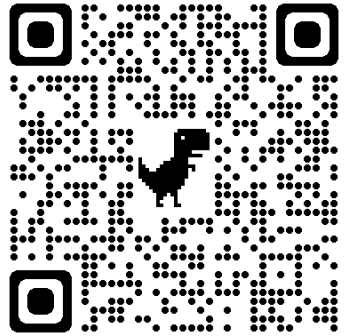


Conditions

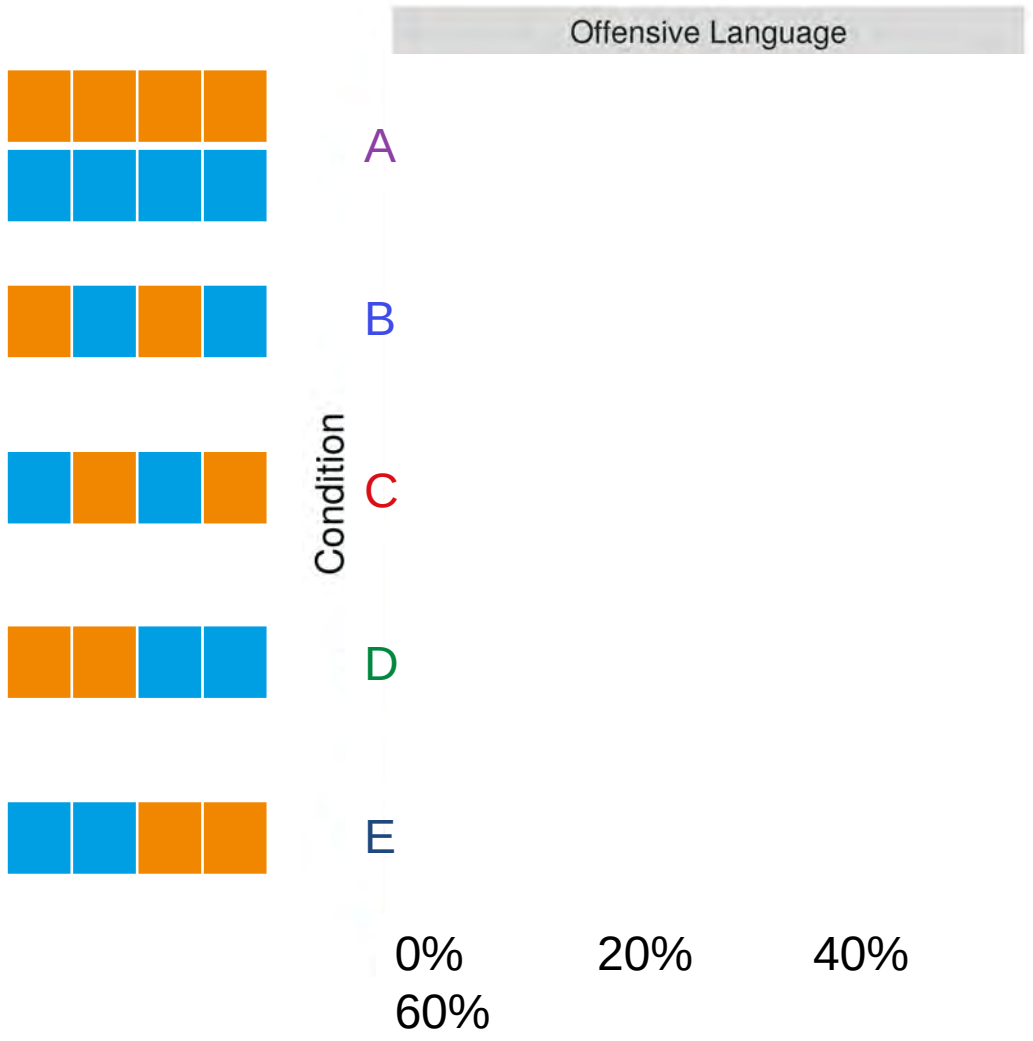
Testing Effects on Quality

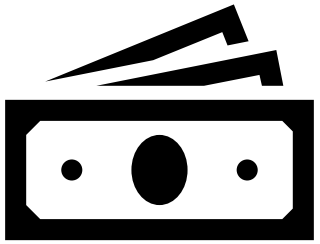
- 3000 tweets (Davidson et al 2017)
- ~900 annotators from Prolific (Nov-Dec 2022)

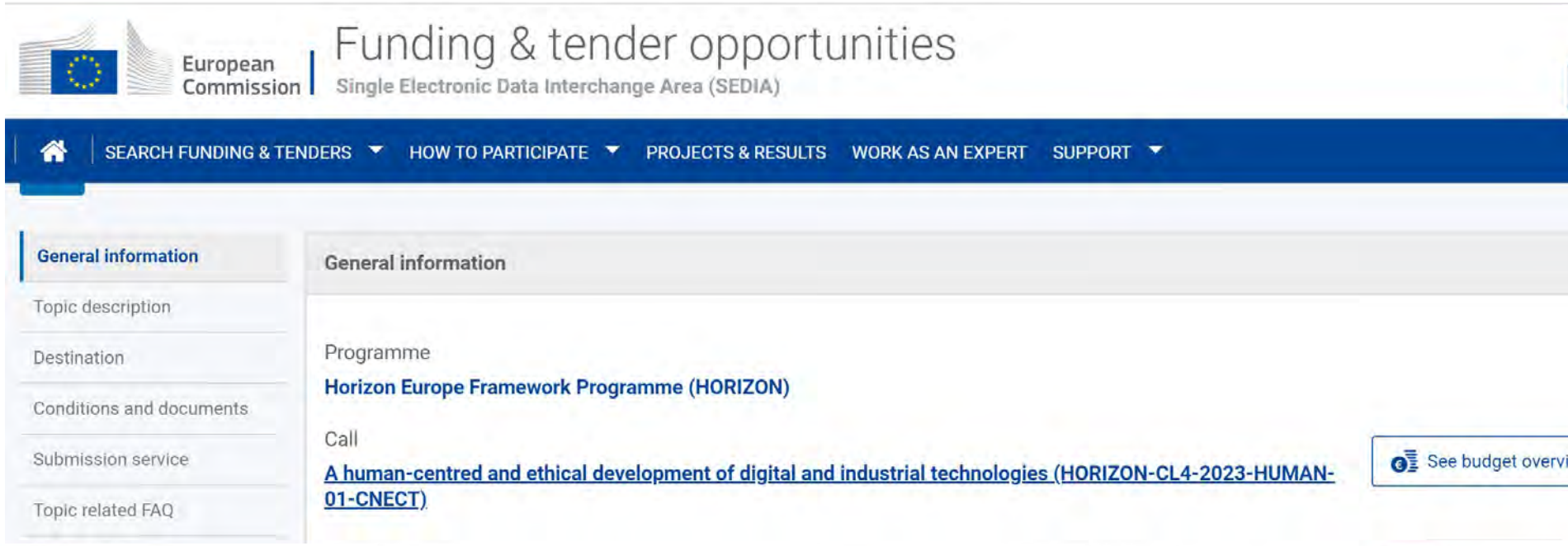
- 50 tweets / annotator
- 3 annotations / tweet - condition
- 15 total annotations / tweet



Order Effects







FUNDING OPPORTUNITIES WITH SPECIAL EMPHASIS ON AI

AI and Society, supported jointly with the Partnership on AI — NSF's directorates for Computer and Information Science and Engineering and Social, Behavioral and Economic Sciences, together with the Partnership on AI, have jointly supported Early-concept Grants for Exploratory Research to understand the social challenges arising from AI technology and enable scientific contributions to overcome them. With increases in the scale and diversity of deployments of AI systems comes the need to better understand AI in the open world, including unforeseen circumstances and social impacts, and to craft approaches to AI that consider these from the start.

PARTNERSHIP ON AI

Fairness, Ethics, Accountability, and Transparency — NSF invites researchers to submit proposals to its core programs that contribute to discovery in research and practice related to fairness, ethics, accountability and transparency in computer and information science and engineering, including AI.

NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon — NSF and Amazon are partnering to jointly support research focused on fairness in AI, with the goal of contributing to trustworthy AI systems that are readily accepted and deployed to tackle grand challenges facing society. Specific topics of interest include, but are not limited to, transparency, explainability, accountability, potential adverse biases and effects, mitigation strategies, validation of fairness, and consideration of inclusivity.

amazon

Real-Time Machine Learning — NSF and the Defense Advanced Research Projects Agency (DARPA) have teamed up to explore high-performance, energy-efficient hardware and machine learning architectures that can learn from a continuous stream of new data in real time. Both agencies issued calls for proposals focused on real-time machine learning and are now offering collaboration opportunities to awardees from both programs throughout the duration of their projects. This partnership is contributing significantly to the foundation for next-generation co-design of algorithms and hardware.

DARPA

DATA AND ADVANCED COMPUTING: A DRIVER OF MODERN AI