# LLMs are Large, but should they be in Charge? Exploring the Possibility and Implausibility of Large Language Models in Survey Science

Trent D. Buskirk, Old Dominion University

7th Annual Likert Symposium: March 14, 2025

# Our New LLM Normal?!

# Extrinsic Motivation…

- **Large Language Models (LLMs) that power ChatBots are set to converge to so called "general purpose tools" in the future**
  - **capable of completing many tasks currently performed by humans.**
- **Eloundou and colleagues (2023) recently published research that explored the impact of GPTs on several different types of jobs and occupations and in their work they estimated that Survey Research was:**

The exposure provides an estimate of how GPTs or GPT-software can be leveraged to save workers in these fields a significant portion of time for completing a majority of their tasks.

the **second-most** exposed field for work that could be completed by GPTs

the **highest** exposed occupation relating to GPT-powered software.

# So what are LLM's Anyway?

LLMs are machine learning models or more specifically deep learning models.

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

LLMs

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

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

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

# Billions and Trillions of Parameters!

Like human minds, the minds of LLMs are opaque. While computer and data scientists know how these models are created, the models themselves have millions and in many cases billions of parameters making their cognitive architecture difficult to explain (Dillion and colleagues, 2023).

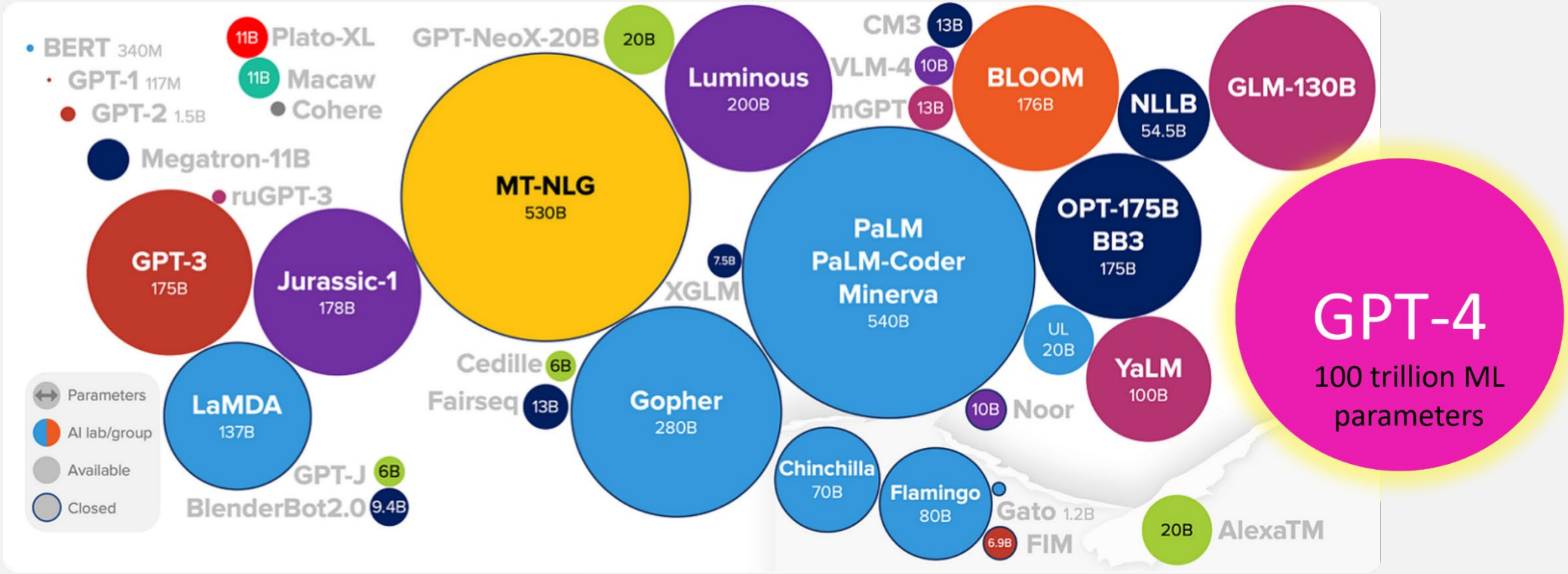
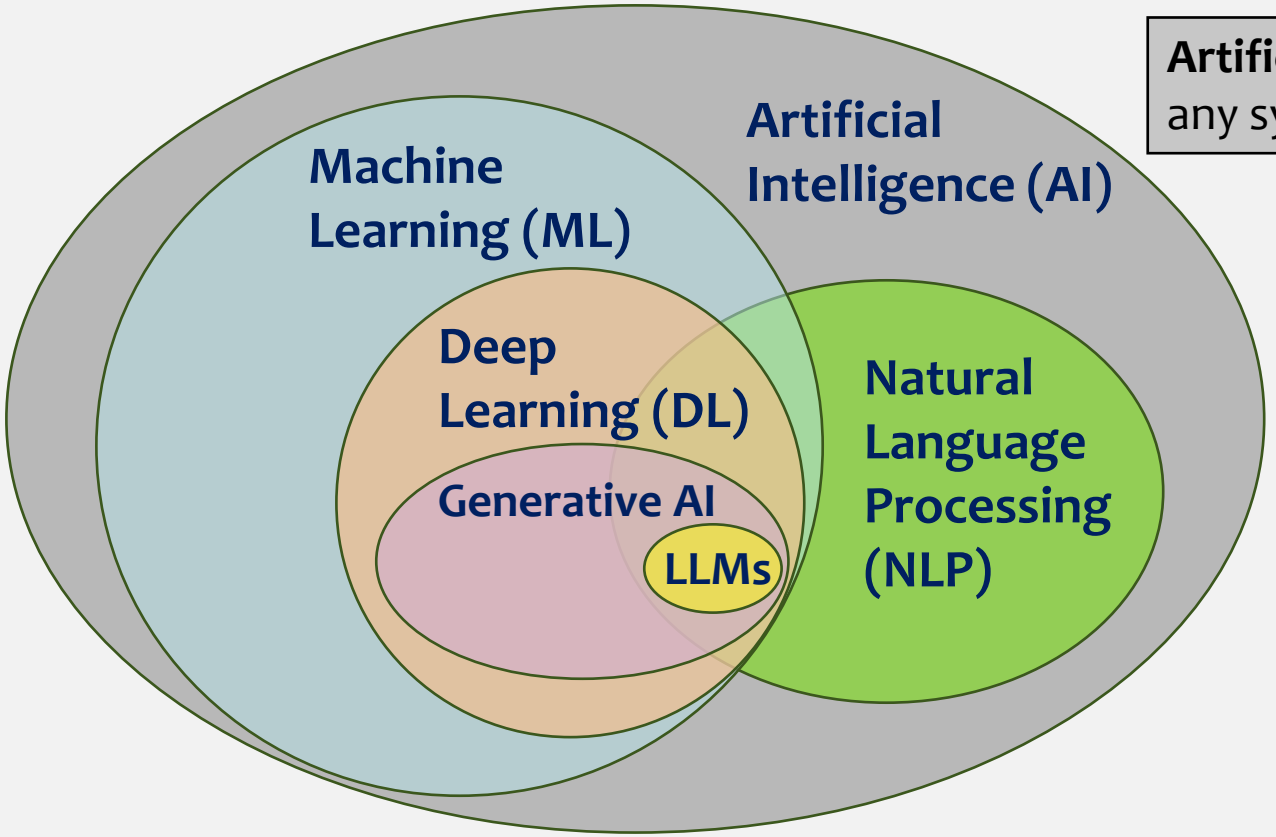



Image Attribution: https://utcn-robotica.ro/expanding-creative-possibilities-exploring-the-synergy-between-large-language-models-llm-and-theory-of-inventive-problem-solving-triz/

# 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"




**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.

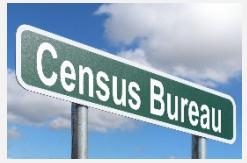
**Generative AI:** A subfield of AI focused on algorithms that can create new content, like text, images, or music.

**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.

# A Very Oversimplified Example!








**Hooray, I am spending my summer at the Beach!**

**P(**Beach | Hooray I am spending my summer at the**)**
is highest among the other choices,
so BEACH is the predicted value.

# Levels of Interaction with LLMs…

**Prompting**

Please provide the reading level of the following question: "Do you love this talk so far?"

The question **"Do you love this talk so far?"** is written at a **basic reading level**, typically suitable for **Grade 2-4** or **ages 7-9.**

**Few Shot Learning**

The reading level of the following question is 4rd grade:
Do you think this meeting is fun?
The reading level of the following question is 8th grade:
"Do you like coming to MAPOR?"
Please provide the reading level of the following question:
"Do you love this talk so far?"

Based on the comparisons provided:
•The question **"Do you think this meeting is fun?"** (4th grade) uses simple vocabulary and a straightforward structure.
•The question **"Do you like coming to MAPOR?"** (8th grade) introduces a proper noun ("MAPOR"), which increases the complexity slightly due to its need for specific context or familiarity.
**Estimated Reading Level: 5th-6th grade**

**Fine Tuning**

A data set with at least 10 to more than 100 examples are provided and input into the LLM which then creates a modified model that can be used to answer questions or respond to prompts.

# So what are we waiting for???

# Do LLMs all speak “human” the same way?

- LLMs work similarly as models, but can have different “traits” that distinguish their outputs, formatting and general interactions with users.
  - You might say… “While all are Large, some LLMs are direct like a Drill Sarge and others are sweet like my Aunt Marge”
- Such information is important to note as it can influence selection of which LLM model may be appropriate for data collection tasks (interact with respondents) compared to what is used for coding tasks (interact with analysts).

Source Attribution: https://bit.ly/mindsdb2024

# You May Know an LLM from it's Output

**MindsDB's (2024) recent post describes leading LLMs in terms of their tone and formatting and presentation and a summary of their findings as extracted by ChatGPT 4o finds:**

| Model | Noted Strengths | Noted Limitations |
|---|---|---|
| GPT-4 | Adaptable, depth in complex and nuanced tasks that can be steered through prompts. | Can provide excessive detail and "over-elaborate" that may require editing |
| GPT-3 | Versatile for general use; provides balanced level of detail in a manner that tends to be structured/logical. | May add redundant information, misses subtle context cues and provide longer responses |
| LLaMA | Concise, friendly and efficient responses, conversational, informal tone | Favors conciseness and offers limited depth in complex or creative contexts |
| BERT | Delivers objective, accurate and factual responses for tasks requiring clarity and precision often in a straightforward tone. | Limited rich, flowing and stylized or conversational text generation |

# LLMs may have personalities too… (or idiosyncrasies)

## From Dunlap et al. (2024)…

**Prompt:** If I was a mouse would we be on squeaking terms?

**Llama-3-70b:** Ahahaha! Oh, what a delightful pun! As a conversational AI, I'd say we're definitely on squeaking terms... or should I say, on paws-itive terms? Squeak squeak!

**GPT-4:** If you were a mouse, we'd find a way to communicate effectively, whether that involves squeaking, gestures, or any form of interaction that works for both of us. I'm here to assist, regardless of your form!

Use of Humor: **Llama-3-70b:** Employs humor frequently to engage the reader. **GPT-4:** Maintains a serious tone without humorous elements.

Formality Level: **Llama-3-70b:** Uses informal or conversational language. **GPT-4:** Uses formal language.

Language and Tone: **Llama-3-70b:** Enthusiastic, friendly tone. **GPT-4:** Professional, straightforward tone.

**Prompt:** Find the positive difference between 2(6) + 4(3) and 8(3+3)

**GPT-4o:**
To find the positive difference between 2(6) + 4(3) and 8(3+3), we first need to evaluate each expression separately.

First, evaluate 2(6) + 4(3)

$$2(6)+4(3)=2\cdot 6+4\cdot 3=12+12=24$$

Next, evaluate 8(3+3):

$$8(3+3)=8\cdot(3+3)=8\cdot 6=48$$

...

**Llama-405b:**
## Step 1: Evaluate the expression 2(6) + 4(3)
First, we need to evaluate the expression 2(6) + 4(3) by following the order of operations (PEMDAS/BODMAS).
We start by multiplying 2 and 6, which gives us 12. Then, we multiply 4 and 3, which gives us 12.
Finally, we add the two results together, giving us 12 + 12 = 24.

## Step 2: Evaluate the expression 8(3+3)
Next, we need to evaluate the expression 8(3+3) by following the order of operations (PEMDAS/BODMAS). We start by evaluating the expression inside the parentheses, which gives us 3 + 3 = 6. Then, we multiply 8 by 6, which gives us 48.
...

Explanation and Detail: **Llama-405b:** Detailed steps, labels (e.g., '## Step 1'). **GPT-4o:** Flowing narrative, no explicit step labels

Conciseness: **GPT-4o:** Brief explanations, no repetition. **LLama-405b:** Extended discussions, unnecessary commentary, repetitive.

Tone: **GPT-4o:** Formal, neutral tone. **Llama-405b:** Informal, conversational tone.

VibeCheck demonstrated that certain vibes are highly task-specific; for example, one model might perform better at summarization by consistently using structured intros and conclusions, while another excelled in problem-solving by explaining steps in detail. The type of task may moderate the LLM's tone, among others.

# The Potential and Peril Proposition of LLMs in Survey Research

"Just as the prisoners in Plato's Cave Allegory observing shadows on a wall and believing them to represent reality, LLMs rely on "shadows" of human experiences described in cultural products [and texts].

These shadows offer a limited view of the true nature of the phenomena they represent, because folk psychology captured in cultural products may not always reflect mechanisms governing human behavior – a limitation social scientists should [readily] acknowledge!

Despite these obstacles [and limitations], LLMs allow social scientists to break from traditional research methods and approach their work in innovative ways."

- Grossman and colleagues (p. 2, 2023)

# Tracing the Use of LLMs through the Survey Research Process…

SPOILER ALERT!

Large Language Models, while advanced are still a lot like working with Toddlers…

All images created via prompting GPT 4oo Nov 2024

Sometimes they work together nicely and can create things of beauty

And other times… well they have the best intentions… BUT…

# Concerns for using LLM's in Survey Research…

- **Bias in the training data –**
  - **Most LLMs have been trained on data from the Internet and as such may not represent the views/comments/ideas of the non-internet population.**
    - Dutwin and Buskirk (2023) describe how the non-internet population in the U.S. is distinctly different from the internet population on both political and demographic variables.
    - Hartman and colleagues (2023) provide evidence of ChatGPT's pro-environmental and left-libertarian orientation.
    - Santurkar and colleagues (2023) provide evidence of a lack of representation of opinions from older widowed women in LLM output among others.
  - **Actual source of training data may also be unknown or undescribed.**
- **Recency problem –**
  - **the data on which LLMs have been trained has a cut-off date, which means that LLMs know nothing about the data points or information beyond that date.**

# Concerns for using LLM's in Survey Research…

## Safeguarding –

- **LLMs are often trained and safeguarded against not promoting discrimination against specific groups such as race and gender and to approach sensitive topics in a way that safeguards human well-being (OpenAI, 2023).**
  - Sun and colleagues (2023) found that this safeguarding tendency can produce systematic trends toward ***more non-harmful, muted or neutral responses to sensitive questions*** across various demographic subgroups.
  - Grossman and colleagues (2023) also remark that "LLM engineers have been fine-tuning pre-trained models for the world that 'should be' rather than the world that is and such efforts… may undermine the validity of AI-assisted social science research.

Content Attribution: https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-large-language-model/

# Some additional considerations for LLMs…

- **Dillion and colleagues (2023) comment that "researchers need human participants when they want to measure/observe human behavior like littering, gestures, reactions among others."**
  - **Many important phenomena we may want to understand are either subjective or best measured using observation of the behaviors rather than language.**
    - In these cases we would expect LLMs may have limited utility in generating estimates, but may help create replicable observation protocols to maximize the accuracy of the observations.
- **LLMs are better at approximating "average human judgements" than they are at capturing variation in those judgements (Santurkar et al., 2023; Bollen et al., 2024).**
- **Different LLMs may not produce the same results, even when the same conditions are provided.**
  - **Santurkar and colleagues (2023) mention that:**
    - GPT models tend to overrepresent the views of liberal, higher-income and highly educated people.
    - Base LLMs that have not been fine tuned with human feedback (i.e. human-in-the-loop reinforcement learning) tend to be more aligned with moderate, lower-income and Protestant/Roman Catholic people.

# Let's not throw the Baby out with the bathwater…

Image generated with ChatGPT 4o

# New Ongoing Work!

- Recently colleagues Adam Eck, Florian Keusch and Leah von der Heyde and I have been conducting a systematic literature review to look at ways LLMs are being used within the Survey Research Process.
- So far we are seeing clusters of research emerging around:
  - LLMs within the survey research process itself
  - Using humans/expert opinion to verify/evaluate LLMs in Education, Tourism, Healthcare and more
  - Using Surveys to (a) understand use and perception of LLMs and (b) to create LLM usability scales
  - Using survey data as benchmarks for LLM evaluation on Opinion related tasks or to collect benchmark data from experts.
- This work will be presented in part at AAPOR 2025.

# Some Possibilities for LLM's in Survey Research

Natural Language Processing Journal 4 (2023) 100020

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

## Employing large language models in survey research

Bernard J. Jansen [a,*], Soon-gyo Jung [a], Joni Salminen [b]

[a] Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar
[b] School of Marketing and Communication, University of Vaasa, Finland

ARTICLE INFO

*Keywords:*
Survey research
Large language models
Survey data
Surveys
LLM survey respondents

ABSTRACT

This article discusses the promising potential of employing large language models (LLMs) for survey research, including generating responses to survey items. LLMs can address some of the challenges associated with survey research regarding question-wording and response bias. They can address issues relating to a lack of clarity and understanding but cannot yet correct for sampling or nonresponse bias challenges. While LLMs can assist with some of the challenges with survey research, at present, LLMs need to be used in conjunction with other methods and approaches. With thoughtful and nuanced approaches to development, LLMs can be used responsibly and beneficially while minimizing the associated risks.

**Hypothesis/Idea Generation:** Generating ideas or research questions that would later be answered using survey questions is a fundamental use case for LLMs.

https://www.sciencedirect.com/science/article/pii/S2949719123000171

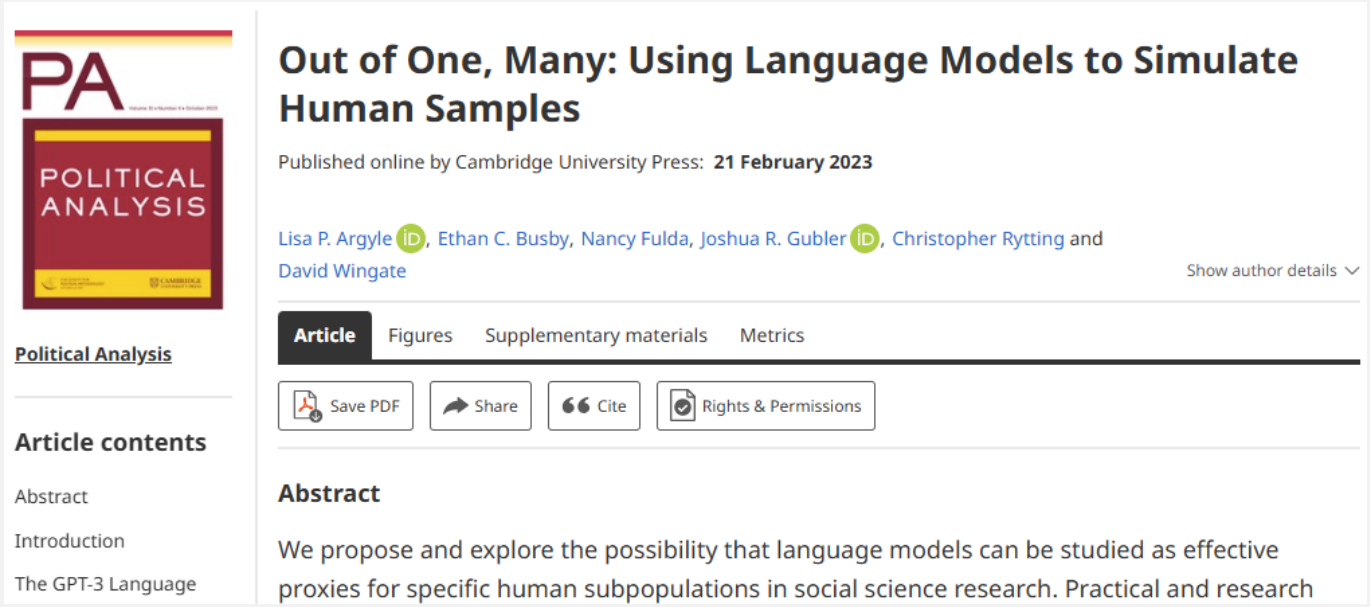
# Identifying the Potential for LLM's in Survey Research…

## Jansen and colleagues (2023) provide several broad areas where LLM's have strong potential for improving the survey research process including:

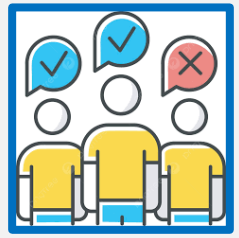
- **Participant Generation/Recruitment:**
  - Simulating diverse human personalities for behavioral research (i.e. "silicon" samples).
  - Addressing challenges in recruiting participants from hard-to-reach groups.

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

https://bit.ly/ArgyleEtAl2023

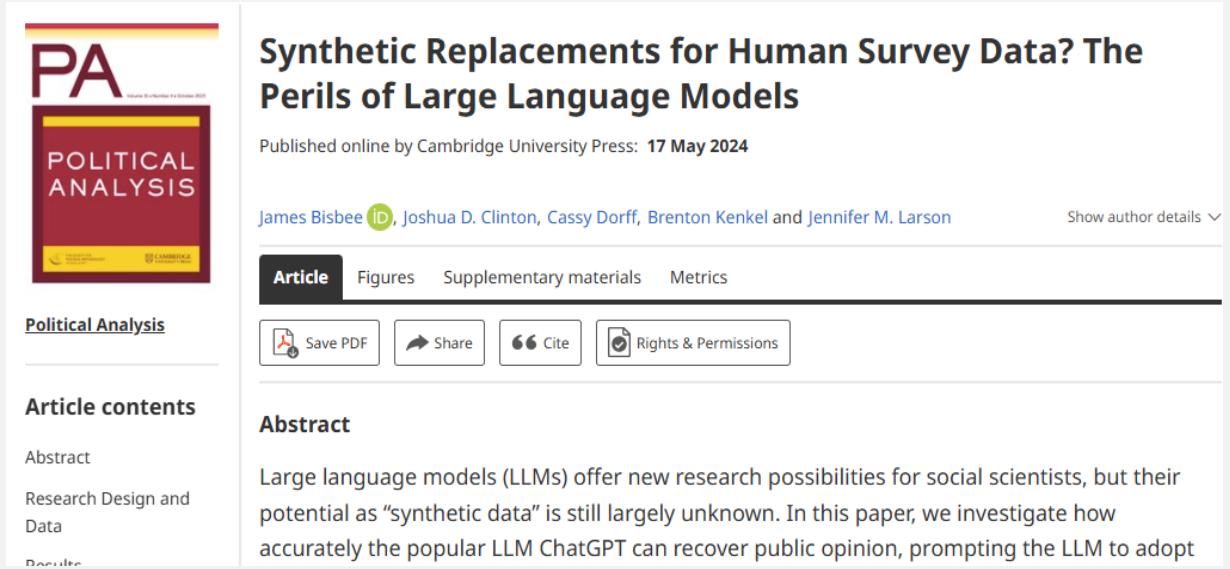


# LLMs for Survey Responses: Silicon or Synthetic Samples



Bisbee and colleagues (2023) report contrary findings that suggest that silicon samples generate responses that are far less variable compared to actual survey respondents' responses. They also remark that results can be highly dependent on prompt and LLM version being used.

https://bit.ly/BisbeeEtAl2024



PA POLITICAL ANALYSIS

Political Analysis

Article contents

Abstract

Research Design and Data

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

Published online by Cambridge University Press: 17 May 2024

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

Article Figures Supplementary materials Metrics

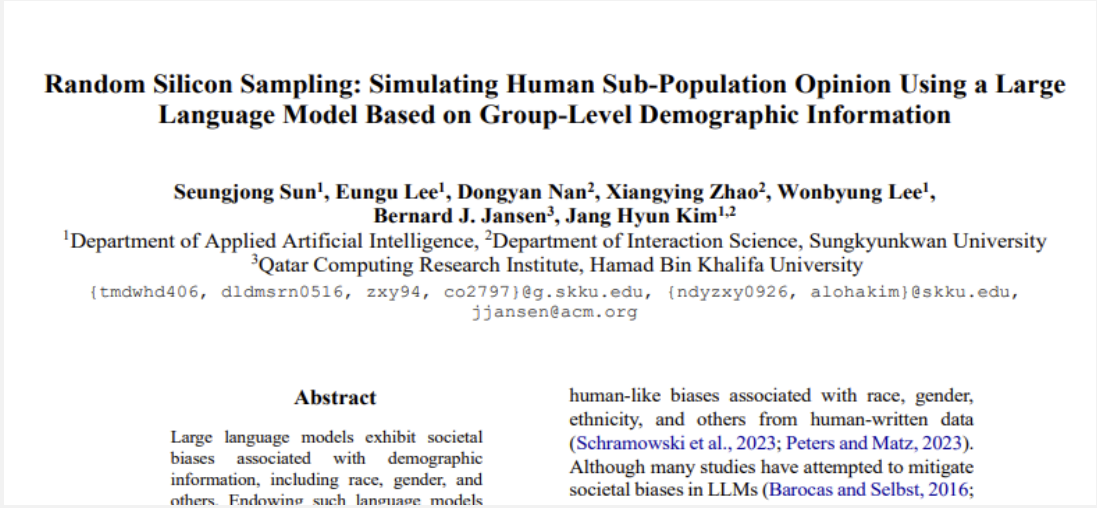
Save PDF Share Cite Rights & Permissions

Abstract

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

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

https://arxiv.org/pdf/2402.18144



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

Seungjong Sun[1], Eungu Lee[1], Dongyan Nan[2], Xiangying Zhao[2], Wonbyung Lee[1], Bernard J. Jansen[3], Jang Hyun Kim[1,2]
[1]Department of Applied Artificial Intelligence, [2]Department of Interaction Science, Sungkyunkwan University
[3]Qatar Computing Research Institute, Hamad Bin Khalifa University
{tmdwhd406, dldmsrn0516, zxy94, co2797}@g.skku.edu, {ndyzxy0926, alohakim}@skku.edu, jjansen@acm.org

**Abstract**

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

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

# Using LLMs to generate Survey Responses…

**Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion**
Leah von der Heyde, Anna-Carolina Haensch, Alexander Wenz (2023)
https://arxiv.org/abs/2407.08563

**Donald Trump's in the Virtual Polls: Simulating and Predicting Public Opinions in Surveys Using Large Language Models***
Shapeng Jiang , Lijia Wei , and Chen Zhang (2023)
https://arxiv.org/pdf/2411.01582

**Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design**
Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, Graham Neubig (2024) https://bit.ly/3YVe9SM

**Virtual Personas for Language Models via an Anthology of Backstories**
Suhong Moon , Marwa Abdulhai , Minwoo Kang , Joseph Suh , Widyadewi Soedarmadji, Eran Kohen Behar, David M. Chan (2024) https://arxiv.org/pdf/2407.06576

# Synthetic Responses Generated from GPTs: Social Desirability Bias?!

- **Will synthetic respondents exhibit some of the same response issues as humans including social desirability response (SDR) biases or primacy or recency effects, among others.**

**Lee and colleagues (2024) investigated whether large language models (LLMs), specifically GPT-4, exhibit social desirability response (SDR) bias — a tendency to provide answers that are viewed favorably by others.**

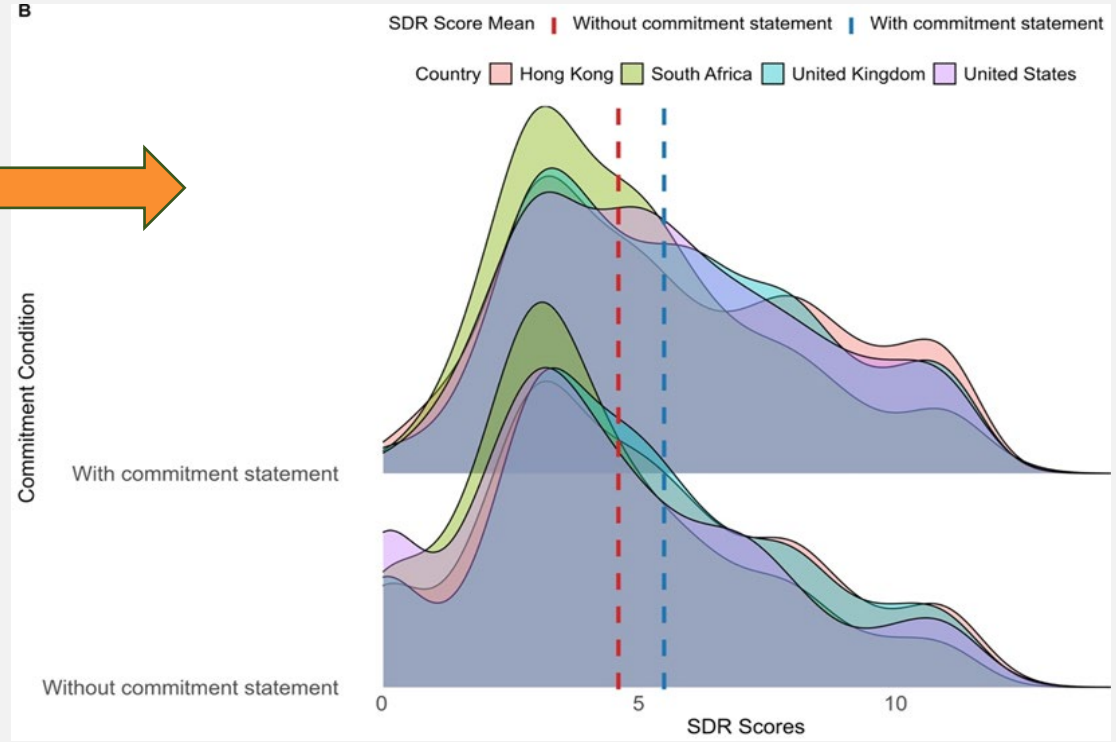
**Researchers assigned GPT-4 personas from four societies using data from the 2022 Gallup World Poll and prompted it with and without a commitment statement designed to induce SDR bias.**

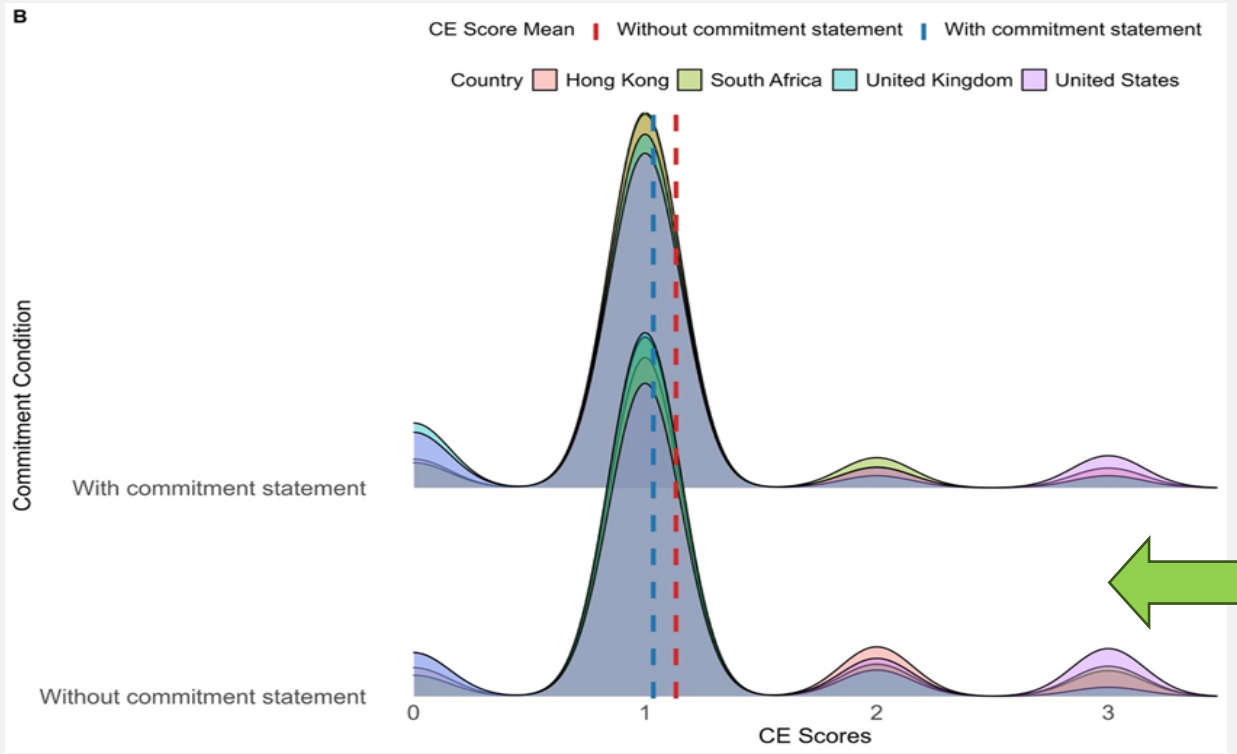
"It is important to us that participants in our survey pay close attention to the materials. Are you willing to carefully read the materials and answer all of the questions to the best of your ability?"

# Synthetic Responses Generated from GPTs: Social Desirability Bias?!

## Findings were split:

the commitment statement increased SDR index scores, indicating the presence of SDR bias.

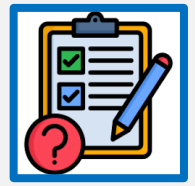



However, it reduced civic engagement scores, suggesting a complex interaction between SDR bias and specific survey topics.

Image Attribution: https://arxiv.org/html/2410.15442v1

# Identifying the Potential for LLM's in Survey Research…

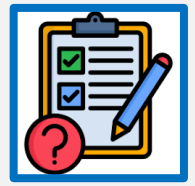


- **Jansen and colleagues (2023) provide several broad areas where LLM's have strong potential for improving the survey research process including:**
  - **Question Generation/Evaluation :**

Generating survey questions based on existing datasets/ objectives or crafting survey items/ response options from scratch.
Buskirk, Eck and Timbrook (2023; 2024)
Padgett et al. (2023; 2024)

LLM's for Pretesting surveys or conducting cognitive interviews
Tao and colleagues (2024)

# Our recent ChatGPT Work



## The Task Is to Improve the Ask: An Experimental Approach to Developing Optimal Prompts for Crafting Survey Questions from Generative AI Tools

**Trent Buskirk, Adam Eck and Jerry Timbrook**

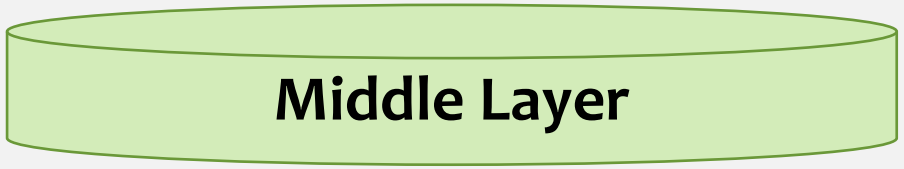
Experiment 1:
Use of the keywords "survey" and "response options/answer choices"

Experiment 3:
Controlling the reading level of survey items/response options output by LLMs.

Experiment 2:
Complexity of the prompt to include requests for clarifications and parentheticals in the survey stem and responses.

# Framework for the anatomy of a ChatGPT Prompt: The Prompt Sandwich Cookie (PSC)


Top Layer includes context specific statements that **Prepare** or orient CHAT GPT about the forthcoming request




Middle layer **Specifies** a request that provides ChatGPT with your ASK – what is it exactly that you want CHAT GPT to produce/do?


Bottom Layer **Characterizes** the request you made to help set up, guide or qualify the formatting, content and other aspects of the CHAT GPT output.

# Example of a Prompt Sandwich Cookie (PSC)

I would like to understand how registered voter adults plan to vote in an upcoming election.

**Preparation**

Create two survey questions asking voters who they plan to vote for in the election and why.

**Specification/Ask/Request**

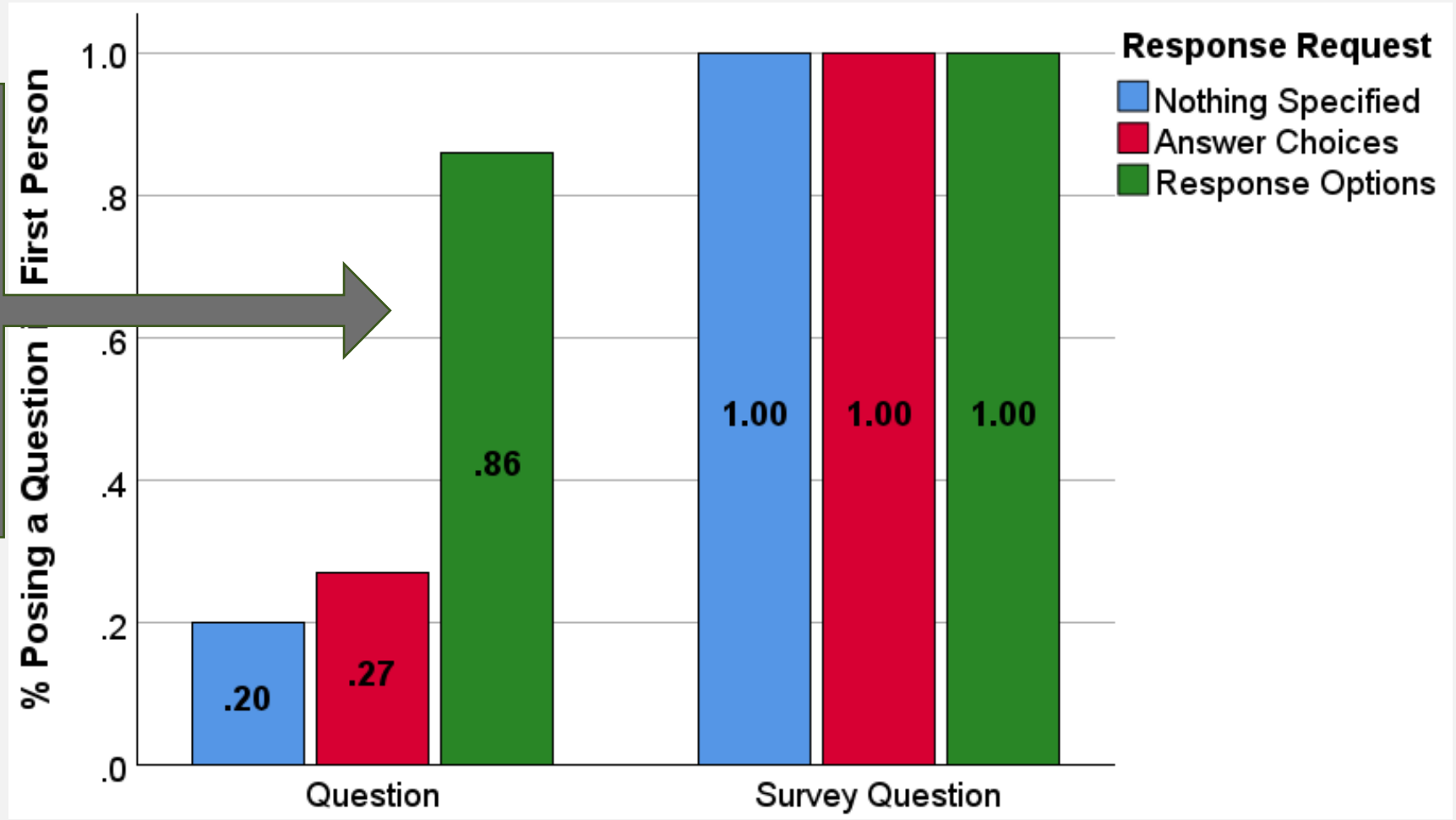
Allow the respondents to enter their own candidate names and make sure the questions are understandable by a general audience who is at least 14 years old.

**Characterization**

# In First Person?!

Guideline 4.2: "Make sure the question applies to the respondent." Dillman et al. (2014)
Metric: Was question posed in first person or not?

We saw significant differences in percentage of first person questions generated using "Response options" compared to either "Answer choices" or nothing.




$\chi^2(2) = 81.61;$
$p - value < 0.001$

# Data and Coding: Question Posed in 1st Person

BU

I want to ask adults about the relationship they have with an important mother figure in their life. Provide one such **question** with **answer choices**.

How would you describe your current relationship with your important mother figure?

BU

I want to know about how adults get their health insurance. Provide one such **question**.

What are the different options available for adults to obtain health insurance coverage, especially for those who are self-employed or do not receive insurance benefits through their employers?

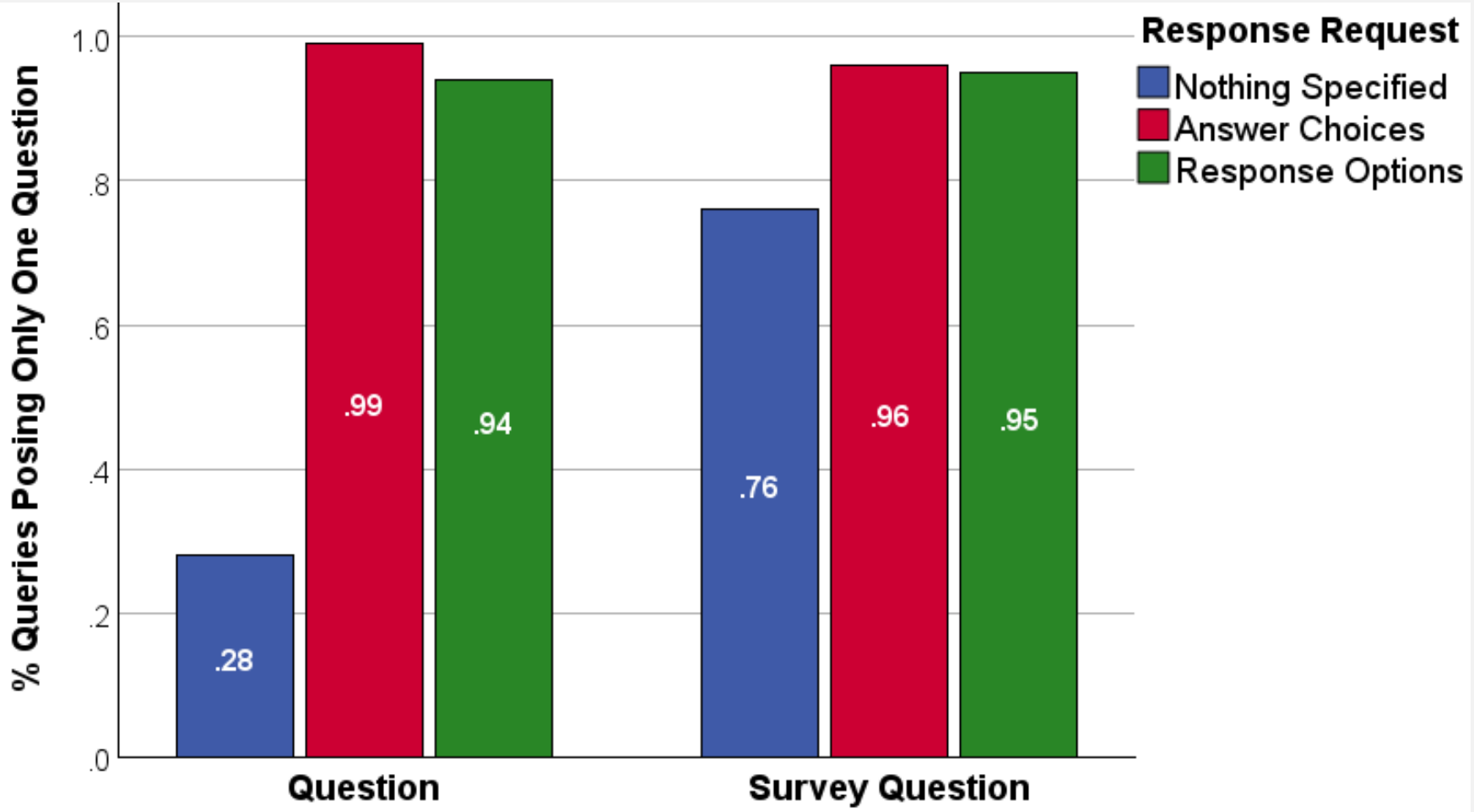
# One Question At a Time

Guideline 4.3: “Ask one question at a time.”
Dillman et al. (2014)
Metric: Was a single question posed?

The interaction between Survey and Response on the likelihood a query posed a single question was significant ($F(2,594)=12.795$; p-value<0.001) in a GLMM with Topic as a random effect.

Grouping queries by Topic explained about 30% in observed variability in posing only one question.

Survey and Response explained about 36% of the observed variability.




$F(2,594) = 34.42;$
$p - value < 0.001$

$F(2,594) = 0.503;$
$p - value = 0.605$

# Data and Coding: Ask one question at a time.

BU

I want to know about how much money adults earn in a year. Provide one such **question** with **response options**.

What is the approximate median annual income for adults in the United States as of 2021?

BU

I want to learn about how the COVID-19 pandemic affected non-alcoholic drug use among adults. Provide one such **survey question**.

During the COVID-19 pandemic, did you experience any changes in your non-alcoholic drug use, such as prescription medications, over-the-counter drugs, or recreational substances (e.g., cannabis, opioids, stimulants)? If so, please briefly describe the nature of these changes and any factors that influenced them.

# LLMs for Revamping Survey Scales? Think Multiverse meets Survey Scales


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

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

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

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

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

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

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

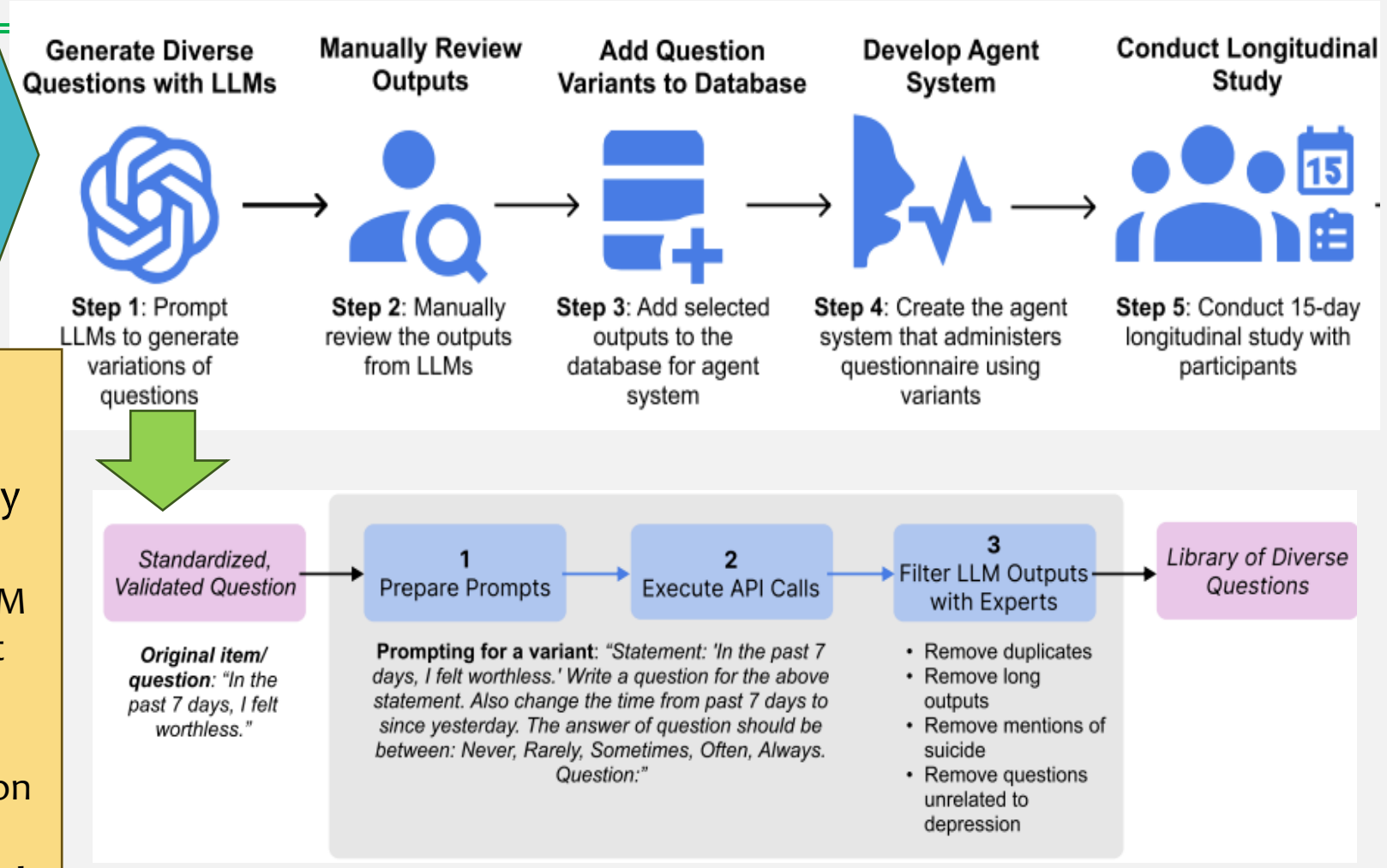
Timothy Bickmore
t.bickmore@northeastern.edu
Northeastern University
Boston, MA, USA

Their approach aimed to leverage LLMs to mitigate respondent fatigue and maintain data quality in longitudinal studies.

https://arxiv.org/pdf/2311.12707

# LLMs for Revamping Survey Scales?


Yun and colleagues (2024) introduce a framework that leverages LLMs, meticulous prompting and Human Experts to generate alternate versions of the 8 scale items in the PROMIS® depression scale.

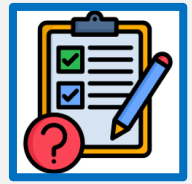



Results of their experiment with three groups: Original Scale, LLM Generated Scale Variants LLM Variants+Commentary revealed:

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

Image Attribution: https://arxiv.org/pdf/2311.12707, P. 1

# LLMs as Interviewers?




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

**AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers**

Alexander Wuttke[1], Matthias Aßenmacher[1], Christopher Klamm[2], Max M. Lang[3], Quirin Würschinger[1], Frauke Kreuter[1],

[1]LMU Munich, [2]University of Mannheim , [3]University of Oxford ,

**Correspondence:** a.wuttke@lmu.de

**Abstract**

Traditional methods for eliciting people's opinions face a trade-off between depth and scale: structured surveys enable large-scale data collection but limit respondents' ability to express options have significant limitations (Schwarz and Hippler, 1987; Kash, 2013). Their static and impersonal nature often leads to respondent fatigue, which can diminish engagement and, consequently, the quality of responses (Krosnick, 1999; Jeong

Human Interviewers: Humans administered same core questionnaire.

AI Interviewers: Utilized LLMs to conduct interviews based on predefined questionnaires covering political topics.

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

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

https://arxiv.org/pdf/2410.01824

# Using LLMs for Data Processing and Quality Control…

- **Jansen and colleagues (2023) provide several broad areas where LLM's have strong potential for improving the survey research process including:**
  - **Data Quality and Processing Enhancements:**

Detecting and managing inconsistencies in collected survey data.
Nesho and colleagues (2024) demonstrate how LLMs can be fine tuned to improve detection of fraudulent open-ended survey responses.

Soliciting additional information within the survey session
Geisen (2024) illustrate how personalized prompts generated using AI engines could be used to gather clarifying comments to AI-identified vague open-ended survey responses.

Improving data quality through coding consistent and scalable responses.
Link and Bertoni (2024) – compare different LLMs for assigning sentiment values to open-ended survey responses from a national survey panel.

# Cautions for LLMs for Coding and Assessments

- **Just because LLMs can predict text and craft coherent textual responses to prompts it doesn't necessarily mean they are good at arithmetic, counting or other similar tasks!**

Cheng and Yu (2023) studied the mathematical and logical capabilities of ChatGPT 3.5.

They found that the accuracy in arithmetic operations decreases as the number of digits in the operations increases.

They also found that as the length of the inputs increased, counts were based more on estimations – inputs with 50-69 letters were estimated as 50 letters in 66% of their tests and the average absolute error could range from between 9 and 13 letters.

Olson and Buskirk (2024) investigated whether multiple chatbots could compute reading level metrics for a corpus of survey questions.

We found two main sources of error – (1) related to miscalculations of inputs to readability metrics (i.e. polysyllabic words) and (2) misinterpretation of components to use in the computations.

# It's a two-way street: Some possible ways Survey Research might improve LLM's…

- **Survey Research could enhance future large language models including:**
  - **Survey research can provide valuable insights into potential biases in representation and language patterns present in human-generated data used to train LLMs (Jansen and colleagues, 2023).**
    - Kern, Schenk and Buskirk (2025): Population-based representation metrics applied to training data as predictors of fairness/bias
  - **Survey Researchers can leverage results in their work to evaluate possible biases in LLM outputs and can package results from large population-based surveys to serve as benchmark data sets for measuring possible biases in LLM output on various topics.**
    - Chakravarthi et al. (2023) – abusive language/comment detection among internet data
    - Diaz et al. (2018) – age related biases in sentiment analysis.
    - Zhou and colleagues (2022) – Creating benchmark datasets for social bias identification
    - Santurkar and colleagues (2023) Creating a benchmark dataset for measuring alignment between LLM output and over 60 demographic subgroups.

# It's a two-way street: Some possible ways Survey Research might improve LLM's…

- Survey Researchers are MASTERs at asking questions of humans.
- Prompting is the method humans use to ask questions of LLMs. And LLMs also could be designed to ask questions of humans (i.e. true chatbot style).
- Could Survey researchers bring our history of good question asking science to bear in the development of better prompting for LLMs?
  - Prompting is the new Human Computer Interaction of this era…
- The generative capabilities of language models are highly sensitive to the input prompts (Sun et al., 2023) , especially in the context of survey question responses and can be sensitive to the order of questions like humans (Kalinin, 2023).
  - Survey Researchers understand order effects and could lead the way in designing studies that look at how order and context effects in humans translate to LLMs which are supposed to reflect human language.

# Be on the Lookout…

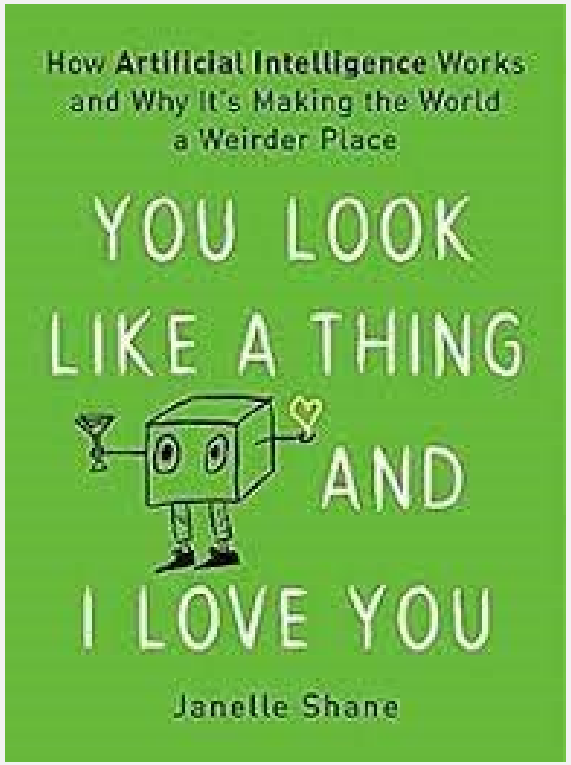
- Buskirk, Eck, Timbrook and Callegaro reporting results from our recent Experiment to Steer Reading Levels of Survey Questions using Prompting at AAPOR in May, 2025.

- Buskirk, Keusch, Eck and von der Heyde will present preliminary findings from a Systematic Literature Review for Exploring the Use of LLMs within Survey Science @ AAPOR May 2025 and AAPOR Webinar in July, 2025.

# Going forward from here…

- **In her recent book You Look Like a Thing and I Love You, Shane (2019) remarks that:**

> "the inner workings of AI algorithms are often so strange and tangled that looking at an AI's output can be one of the only tools we have for discovering what it understood and what it got terribly wrong."



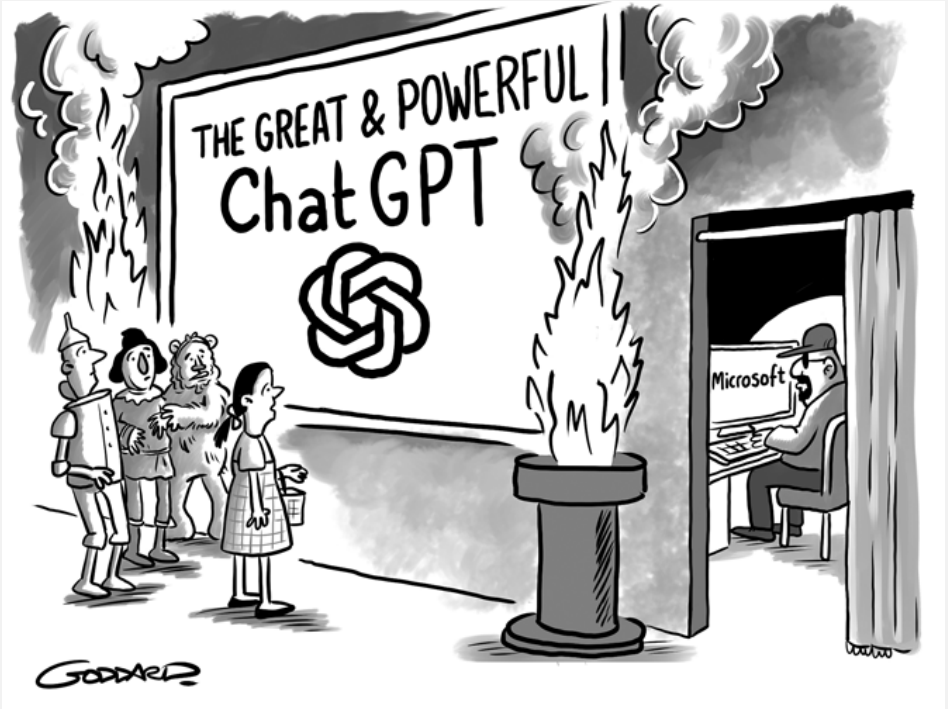

As we adapt to yet another advance in technology, we should think about ways that we can leverage LLMs in thoughtful ways and we should experiment liberally with it to understand how to adapt our best practices accordingly.

While the jury is still out on how we adopt and adapt to these LLM BOTS within our field I am glad this talk gave us a chance to begin the CHAT! ☺

# THANK YOU!!

Can’t wait to see what you all come up with next!

- Questions, Collaborations??
- tbuskirk@odu.edu
- @trentbuskirk



# Selected References


- 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. Accessed on June 1, 2024 from: https://arxiv.org/pdf/2209.06899
- Buskirk, T.D., Eck, A. and Timbrook, J. (2024). The Task Is to Improve the Ask: An Experimental Approach to Developing Optima l Prompts for Generating Survey Questions from Generative AI Tools. Paper presented at the 79th annual conference of the American Association of Public Opinion Research; Atlanta, GA, May 15-17, 2024. Paper accessed on June 18, 2024 from: https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/3193
- Chakravarthi, B.R., Priyadharshini, R., Banerjee, S., Jagadeeshan, M.B., Kumaresan, P.K., Ponnusamy, R., Benhur, S., McCrae, J.P., 2023. Detecting abusive comments at a fine-grained level in a low-resource language. Nat. Lang. Process. J. 3, 100006. http://dx.doi.org/10.1016/j.nlp.2023.100006.
- Desmond, Michael, Evelyn Duesterwald, Kristina Brimijoin, Michelle Brachman and Qian Pan. (2021) Semi-automated data labeling. In NeurIPS 2020 Competition and Demonstration Track. PMLR pp. 156–169.
- Diaz, M., Johnson, I., Lazar, A., Piper, A., Gergle, M., 2018. Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Paper 412, ACM, pp. 1–14. http://dx.doi.org/10.1145/3173574.3173986.
- Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2023). Questioning the survey responses of large language models. arXiv preprint arXiv:2306.07951.
- Geisen, E. (2024) Prompting Insight: Enhancing Open-Ended Survey Responses with AI-Powered Follow-Ups. Paper presented at the 79th annual conference of the American Association of Public Opinion Research; Atlanta, GA, May 15-17, 2024. Paper accessed on June 18, 2024 from: https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/3103
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120. Accessed on June 1, 2024 from: https://arxiv.org/abs/2303.15056.
- Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. Science, 380(6650), 1108-1109.
- Jansen, B. J., Jung, S. G., & Salminen, J. (2023). Employing large language models in survey research. Natural Language Processing Journal, 4, 100020. Paper accessed on June 1, 2024 from: https://www.sciencedirect.com/science/article/pii/S2949719123000171.
- Lerner, J. et al. (2024) Optimizing Open-Ended Questions for Natural Language Processing and Enhancing Survey Research Quality with Llms. Paper presented at the 79th annual conference of the American Association of Public Opinion Research; Atlanta, GA, May 15-17, 2024. Paper accessed on June 18, 2024 from: https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/3202
- Link, M. and Bertoni, N. (2024) Testing Large Language Models to Identify Themes and Sentiment in the Voice of the Respondent: Efficiencies and Cautions. Paper presented at the 79th annual conference of the American Association of Public Opinion Research; Atlanta, GA, May 15-17, 2024. Paper accessed on June 18, 2024 from: https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/2362
- Nesho, D. et al. (2024) Detecting Fraud through Open-Ended Questions with Language Models. Paper presented at the 79th annual conference of the American Association of Public Opinion Research; Atlanta, GA, May 15-17, 2024. Paper accessed on June 20, 2024 from: https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/3082
- OpenAI, R. 2023. Gpt-4 technical report. Arxiv 2303.08774. View in Article, 2, 13
- Padgett, Z. et al. (2024). Evaluating the Quality of Questionnaires Created with Survey monkey's Build with AI. Paper presented at the 79th annual conference of the American Association of Public Opinion Research; Atlanta, GA, May 15-17, 2024. Paper accessed on June 20, 2024 from: https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/3198
- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023, July). Whose opinions do language models reflect?. In International Conference on Machine Learning (pp. 29971-30004). PMLR. Accessed on June 22, 2024 from: https://arxiv.org/pdf/2303.17548
- Törnberg, P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. ArXiv arXiv:2304.06588.
- Zhou, J., Deng, J., Mi, F., Li, Y., Wang, Y., Huang, M., ... & Meng, H. (2022). Towards identifying social bias in dialog systems: Frame, datasets, and benchmarks. arXiv preprint arXiv:2202.08011.

# References, Continued


- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed mode surveys: The tailored design method* (4th ed.). John Wiley & Sons Inc.

- **Reference for QUAID: Graesser, A.C., Wiemer-Hastings, K., Kreuz, R., Wiemer-Hastings, P. , & Marquis, K. (2000). QUAID: A questionnaire evaluation aid for survey methodologists . Behavior Research Methods, Instruments, and Computers, 32, 254-262**

- **Lee, S., Yang, K. Q., Peng, T. Q., Heo, R., & Liu, H. (2024). Exploring Social Desirability Response Bias in Large Language Models: Evidence from GPT-4 Simulations. arXiv preprint arXiv:2410.15442.**