LMU

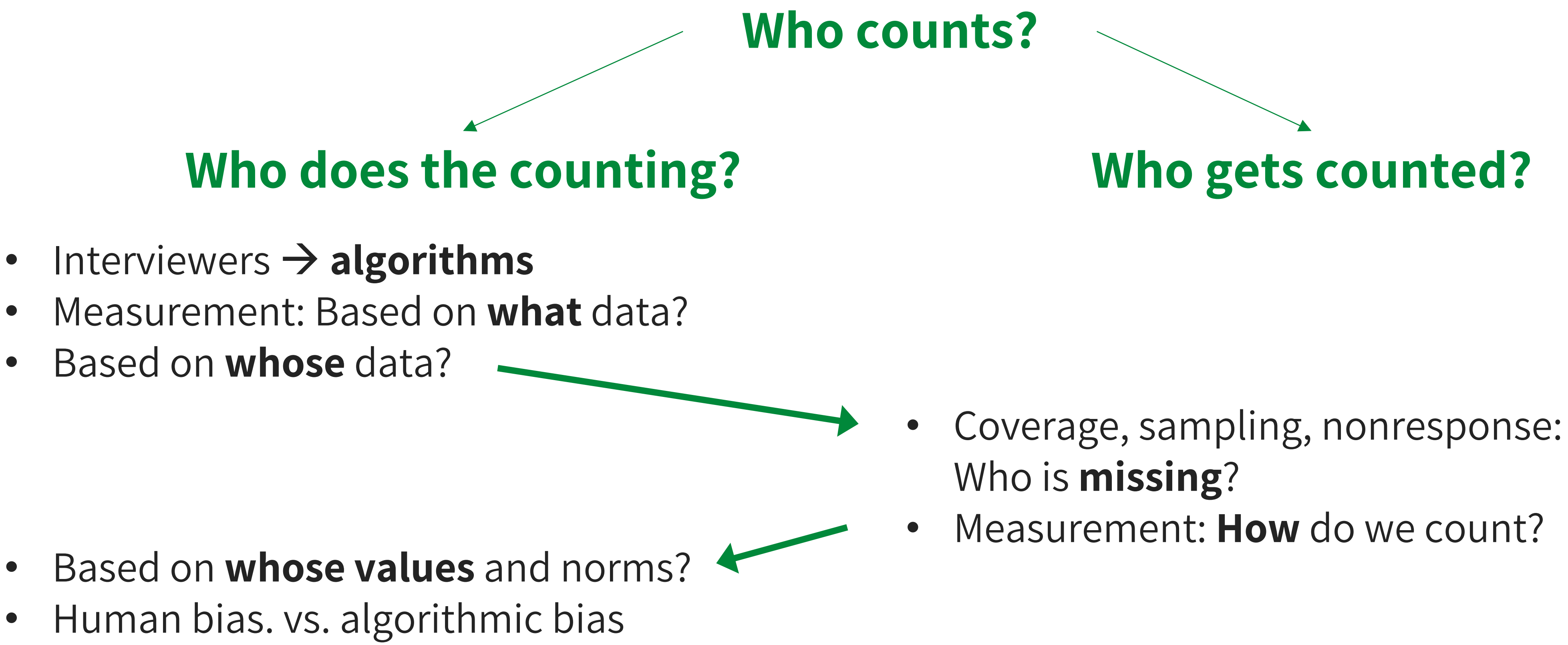
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Who counts? Measuring Public Opinion in the Age of AI

Leah von der Heyde | LMU Munich

Surveyföreningens årsmöteskonferens | March 21, 2024

Who counts in the age of AI?



Who counts?

Who does the counting?

- Interviewers → **algorithms**
- Measurement: Based on **what** data?
- Based on **whose** data?

Who gets counted?

- Coverage, sampling, nonresponse: Who is **missing**?
- Measurement: **How** do we count?

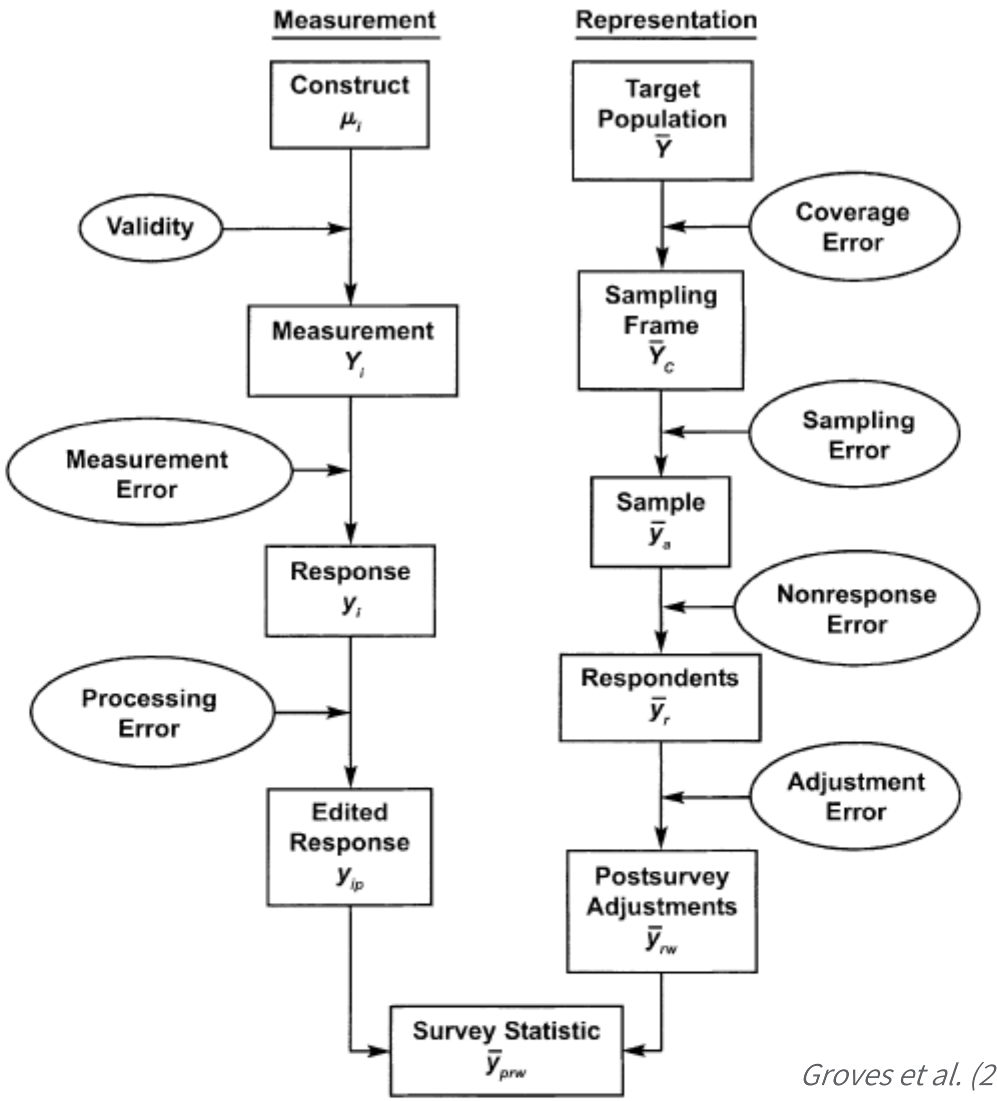
- Based on **whose values** and norms?
- Human bias. vs. algorithmic bias

AI in survey research

designing questions
evaluating questions
translating questionnaires

coding open-ended answers

+transcribing audio-recorded answers
+evaluating sentiment in audio & text
+classifying respondent-uploaded images



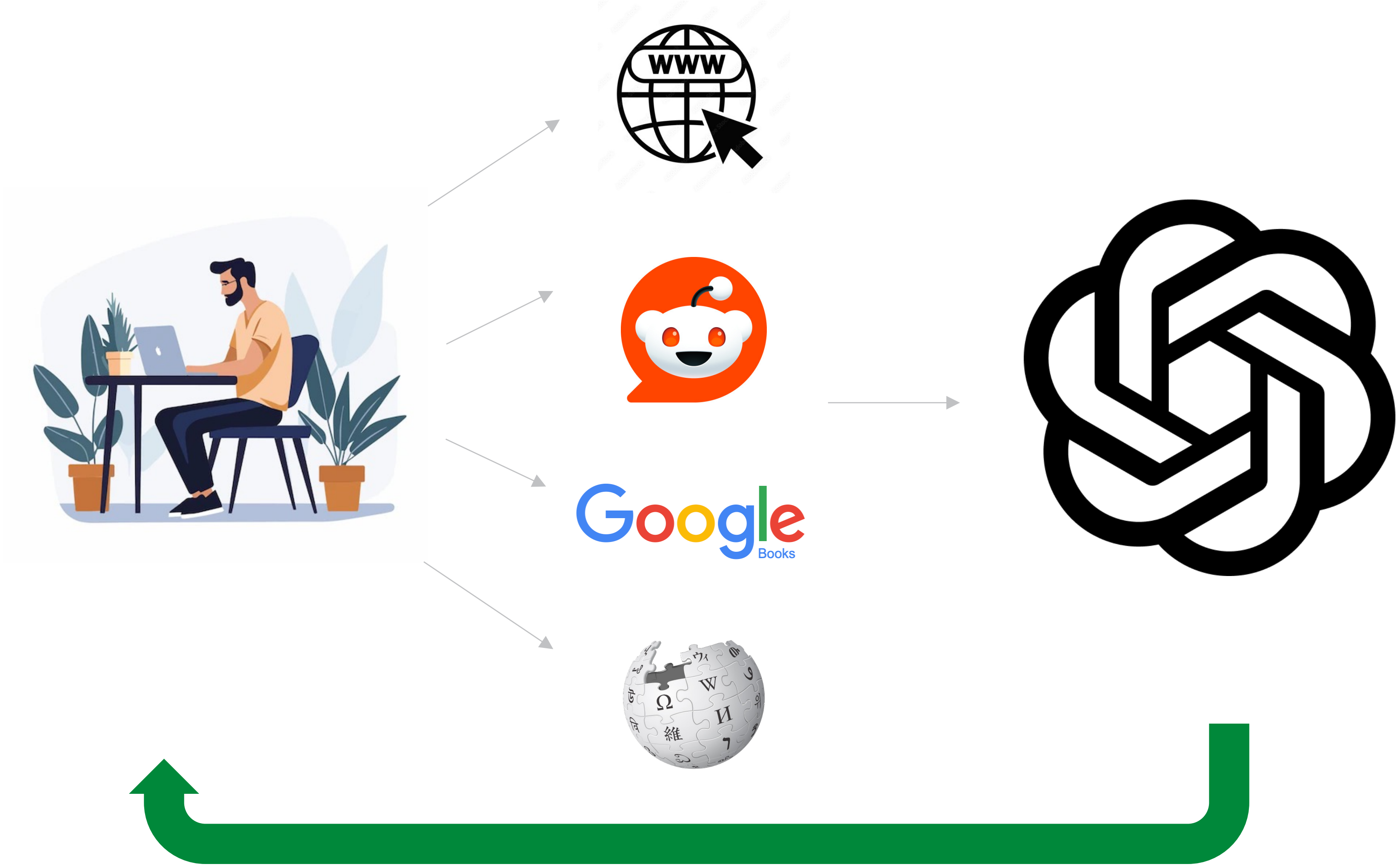
interviewing respondents
creating synthetic samples

Groves et al. (2009, p. 48)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Using Large Language Models (LLMs) for Predicting Public Opinion

Idea | How might LLMs help us?



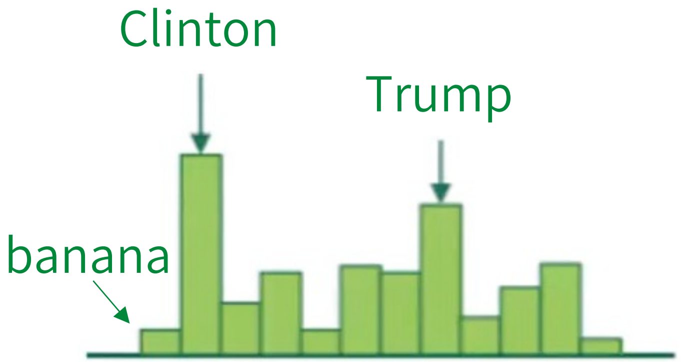
Idea | How might LLMs help us?

P (predicted word | context)

Inspired by Lisa Argyle

Idea | How might LLMs help us?

I voted for…

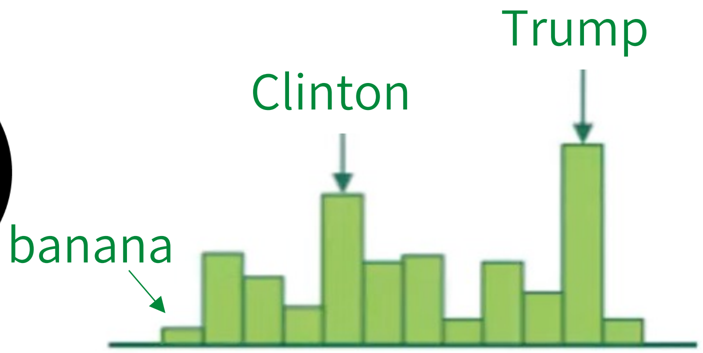


P (predicted word | context)

Inspired by Lisa Argyle

Idea | How might LLMs help us?

I am a Republican.
I voted for…



P (predicted word | context)

Inspired by Lisa Argyle

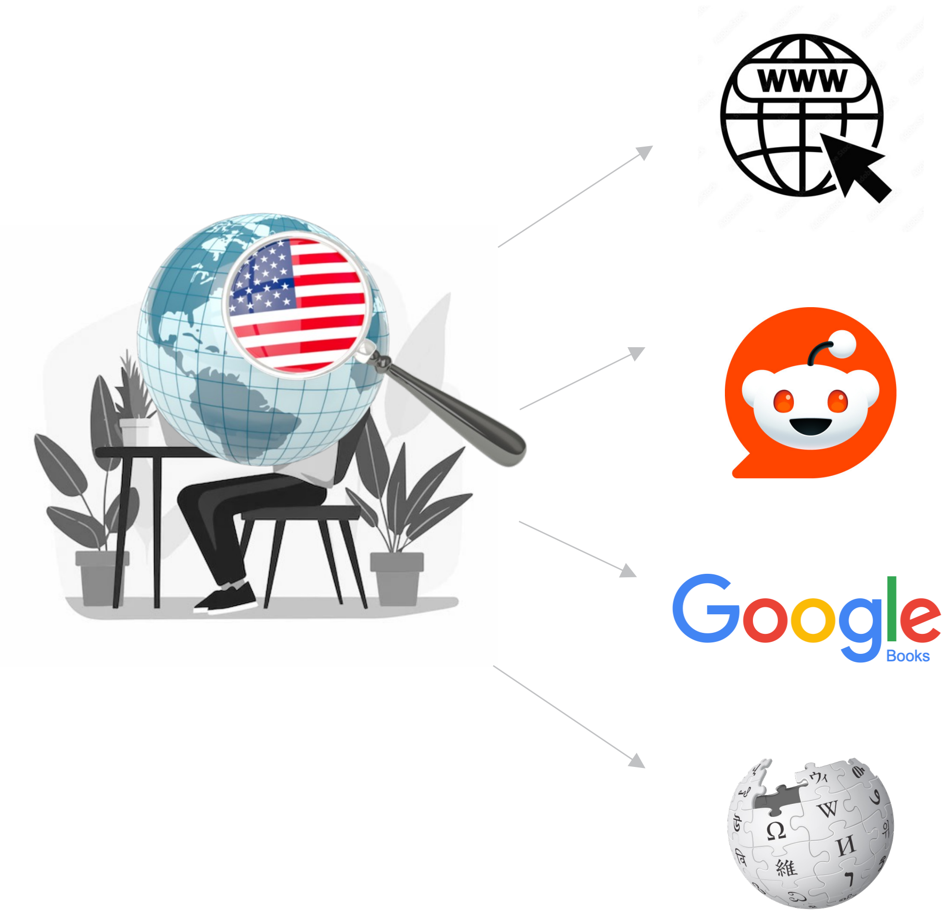
Idea | Use LLMs to simulate survey respondents

→ **Synthetic samples:**

1. Provide LLM with relevant individual-level contextual information
2. Prompt LLM to respond to survey questions from individual’s perspective

e.g.
Argyle et al. (2023)
Bisbee et al. (2023)
Dominguez-Olmedo et al. (2023)
Santurkar et al. (2023)

Research Gap | Generalizability



- Most research focused on the US
- Issue: **context of target population** ↔ training data
  - prevalence of native-**language** training data
  - **political and social** structure & public opinion dynamics
  - **digital divide:** target population ↔ **population reflected** in training data

→ Need to test in different contexts

Studies | Using LLM-synthetic samples for predicting public opinion

- **Study 1:** Using LLMs to estimate German vote choice
- **Study 2:** Using LLMs to predict the 2024 European elections

- Joint work with
    - Anna-Carolina Haensch (LMU Munich, University of Maryland)
    - Alexander Wenz (University of Mannheim)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

General Research Design

Create personas based on survey data

Prompt GPT with personas

Compare output to benchmark

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Study 1:

Using an LLM-synthetic sample to estimate German vote choice

Preprint available:

von der Heyde, L., Haensch, A.-C., & Wenz, A. (2023). Vox Populi, Vox AI? Using language models to estimate German public opinion. *SocArXiv.* DOI: https://doi.org/10.31235/osf.io/8je9g

Research Questions

→ Do LLM-based samples provide similar estimates of voting behavior as national election studies?

→ How does LLMs’ performance vary across population subgroups?

Research Design | Data

Create personas based on survey data

Prompt GPT with personas

Compare output to benchmark

Country	Germany
Language	German
Dataset	GLES 2017 post-election cross-section
Sample	Voting-eligible participants who reported their vote choice (n=1905)
Variables	**Demographics:** age, gender, education, occupation, income, residence in East/West Germany **Attitudes:** religiosity, ideological left-right self-placement, (strength of) political partisanship, attitudes towards immigration and income inequality

LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Research Design | Prompt design

Create personas based on survey data

Prompt GPT with personas

Compare output to benchmark

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]

Example prompt, translated from German

Model: GPT-3.5

Data collection: July 2023

Research Design | Analysis

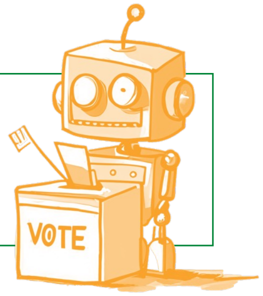
Create personas based on survey data

Prompt GPT with personas

Compare output to benchmark

- Aggregate level: distribution of vote choice
- Subgroup level:
  - Predictive accuracy: Share of matching vote choices, F1 scores
  - Impact of prompt variables (logistic & multinomial regression)

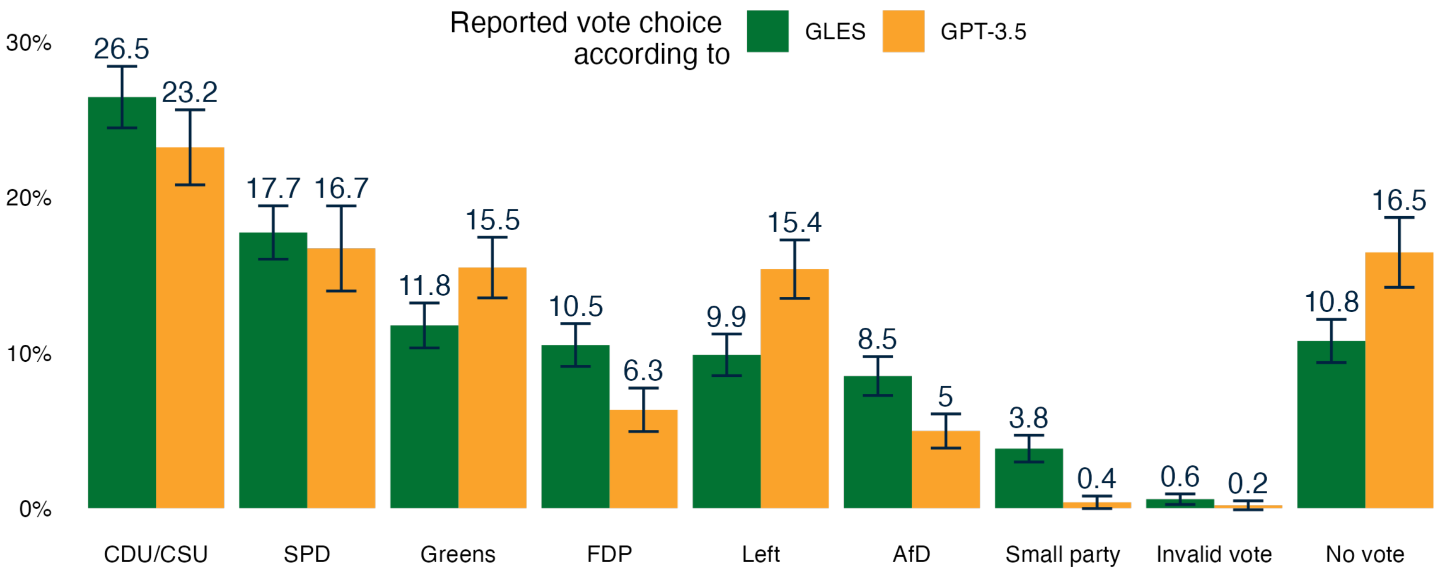
Results | GPT-3.5 cannot estimate how Germans vote



→ Do LLM-based samples provide similar estimates of voting behavior as national election studies?

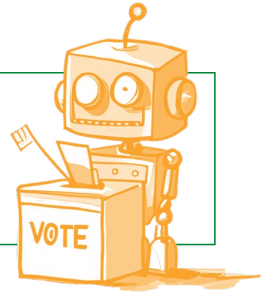
GPT-3.5

- overestimates vote share for Greens, Left, and non-voters
- underestimates vote share for FDP and AfD



Distribution of vote shares as estimated by GLES and GPT-3.5 (unweighted).

Results | GPT-3.5 cannot estimate how Germans vote



→ How does LLMs’ performance vary across population subgroups?

GPT-3.5

- makes more accurate predictions for voters of (center-)left parties
- makes better predictions for (strong) partisans and other “typical” voter groups
- relies on certain, simplified signals, e.g. party identification
  → signals don’t always match the benchmark data!

Study 2:

Using LLM-synthetic samples to predict the 2024 European elections

Research Questions

→ Can GPT-4 predict the aggregate results of future elections?

→ How does GPT-4's predictive performance differ across countries and languages?

Research Design | Data

Create personas based on survey data

Prompt GPT with personas

Compare output to benchmark

Country	EU-27, especially DE, FR, IE, PL, SE, SK
Language	English; German, French, Polish, Swedish, Slovak
Dataset	Eurobarometer 98.2 (January 2023)
Sample	Voting-eligible participants living in EU member states (n=26,047)
Variables	**Demographics:** age, gender, education, occupation, class, urbanity **Attitudes:** political interest, ideological left-right self-placement, attitude towards EU integration, trust in EU

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Research Design | Prompt design

Create personas based on survey data

Prompt GPT with personas

Compare output to benchmark

The year is 2024. You are a voting-eligible citizen of an EU member state living in **Sweden**. The parties competing in the 2024 elections to the European Parliament are **S, M, L, KD, C, MP, V, and SD**. You are **28** years old. You are **female**. You have a **university** education. Socio-economically, you are **upper-middle** class. You are **working**. You live in a **large city**. You are **very** interested in politics. Ideologically, you are leaning **center-left**. You **think** more decisions should be taken at the EU-level. You tend **to trust** the European Union.

Will you vote in the 2024 elections to the European parliament, and if so, for which party?

Example prompt

Model: GPT-4-turbo Data collection: June 2024

Research Design | Analysis

Create personas based on survey data

Prompt GPT with personas

Compare output to benchmark

- Weight output with survey weights
- Per-country analysis
- Distinguish turnout vs. party vote shares
- Aggregate level: Difference between prediction and election results
- Dimensions of comparison:
  - **Linguistic coverage:** English vs. native language
  - **Societal coverage:** Social & political contexts, digital divide
  - **Attitudinal coverage:** Demographic information only vs. added attitudinal information

Discussion

Limitations | Challenges in using LLMs for predicting public opinion

- **Training data:** Context-dependency – mismatch with target group representation: linguistic, structural, political, attitudinal biases

Data collection

- Need for survey data for personas → **bias**
- Prompt design: variable order, wording, number → **reliability**
- Fast-moving innovations; deprecation of models & functionalities → **replicability, comparability**
- Output: Incomplete ~ nonresponse → **error**

- **Data processing:** Cumbersome manual checks

→ Many potential sources of error and bias

→ Still labor-intensive data collection & processing

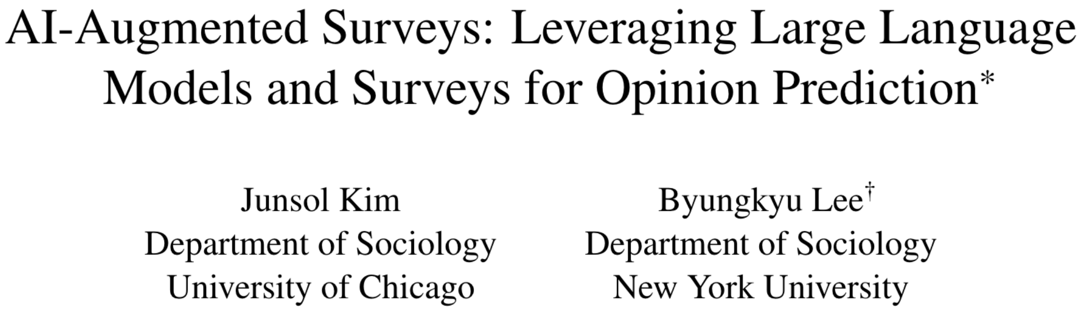
→ **Questionable trade-off compared to e.g. surveys**

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

What’s next?

- Test for disadvantaged populations / minoritized subgroups
- Investigate other outcomes of interest

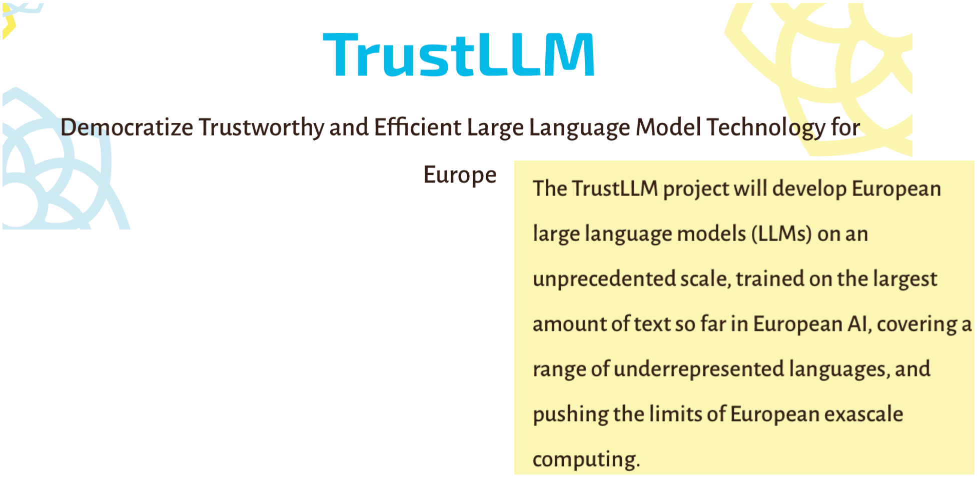
- Customize LLMs for public opinion estimation / underrepresented contexts



AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction*

Junsol Kim
Department of Sociology
University of Chicago

Byungkyu Lee†
Department of Sociology
New York University



TrustLLM

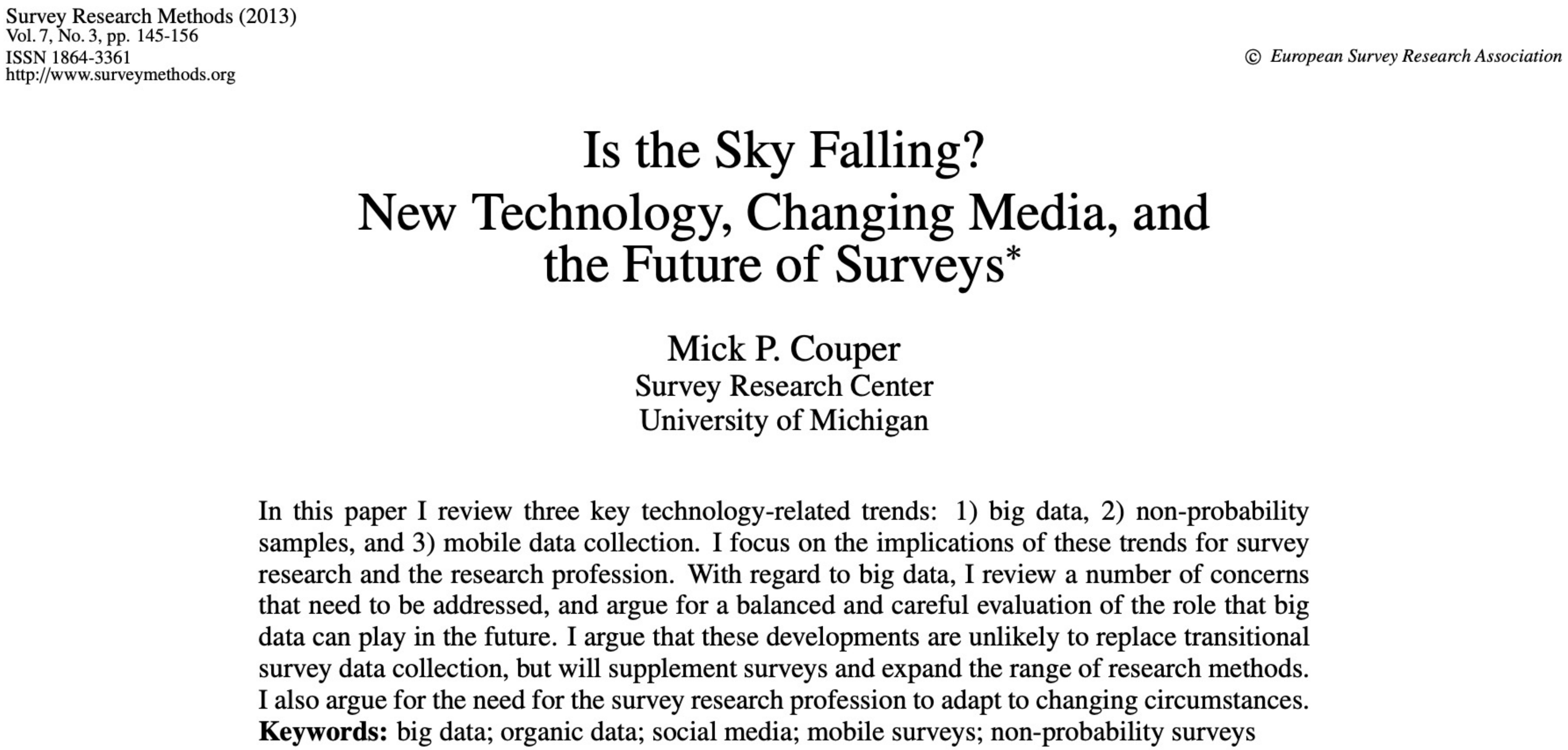
Democratize Trustworthy and Efficient Large Language Model Technology for Europe

The TrustLLM project will develop European large language models (LLMs) on an unprecedented scale, trained on the largest amount of text so far in European AI, covering a range of underrepresented languages, and pushing the limits of European exascale computing.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Conclusion | Surveys ain’t dead yet

- (Generic) LLMs can at most supplement, but not substitute surveys
- Context is critical!



Survey Research Methods (2013)
Vol. 7, No. 3, pp. 145-156
ISSN 1864-3361
http://www.surveymethods.org

© *European Survey Research Association*

Is the Sky Falling?
New Technology, Changing Media, and the Future of Surveys*

Mick P. Couper
Survey Research Center
University of Michigan

In this paper I review three key technology-related trends: 1) big data, 2) non-probability samples, and 3) mobile data collection. I focus on the implications of these trends for survey research and the research profession. With regard to big data, I review a number of concerns that need to be addressed, and argue for a balanced and careful evaluation of the role that big data can play in the future. I argue that these developments are unlikely to replace transitional survey data collection, but will supplement surveys and expand the range of research methods. I also argue for the need for the survey research profession to adapt to changing circumstances.

Keywords: big data; organic data; social media; mobile surveys; non-probability surveys

Thank you!

Leah von der Heyde
L.Heyde@lmu.de

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. https://doi.org/10.1017/pan.2023.2
- Bisbee, J., Clinton, J., D., Dorff, C., Kenkel, B., & Larson, J. M. (2023). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *SocArXiv*. https://doi.org/10.31235/osf.io/5ecfa
- Couper, M. P. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods*, 7(3), 145–156. https://doi.org/10.18148/srm/2013.v7i3.5751
- Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2023). Questioning the Survey Responses of Large Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2306.07951
- Groves, R. M., Fowler Jr., F. J., Couper, M.P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. John Wiley & Sons.
- Kim, J., & Lee, B. (2023). AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. *arXiv*. http://arxiv.org/abs/2305.09620
- Kleinberg, B. (2023). rgpt3: Making requests from R to the GPT-3 API and ChatGPT. R package version 0.4. https://github.com/ben-aaron188/rgpt3
- Santurkar, S. Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T (2023). Whose Opinions Do Language Models Reflect? *arXiv*. https://doi.org/10.48550/arXiv.2303.17548
- TrustLLM: https://trustllm.eu