LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

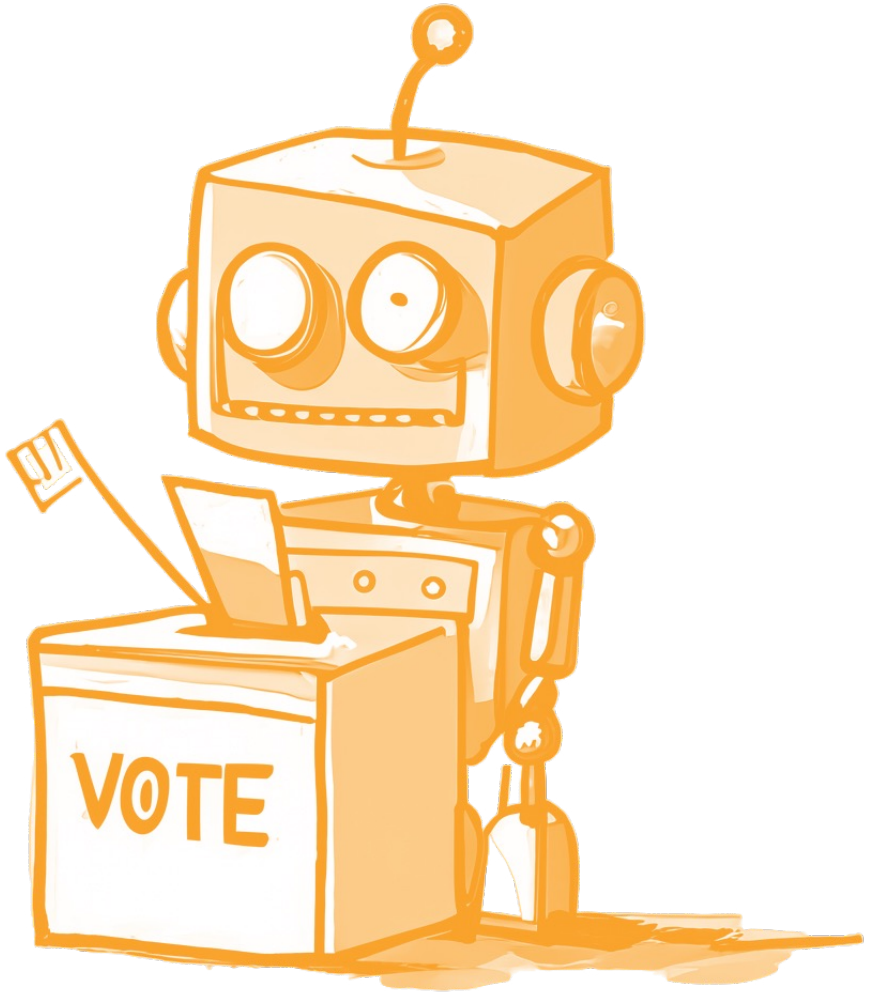
# Can Large Language Models Estimate How People Vote?

## Evidence from Germany

**Leah von der Heyde** | LMU Munich

*work with*
Anna-Carolina Haensch | LMU Munich, U. of Maryland
Alexander Wenz | University of Mannheim

**WEB DATA OPP | March 19, 2024**



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

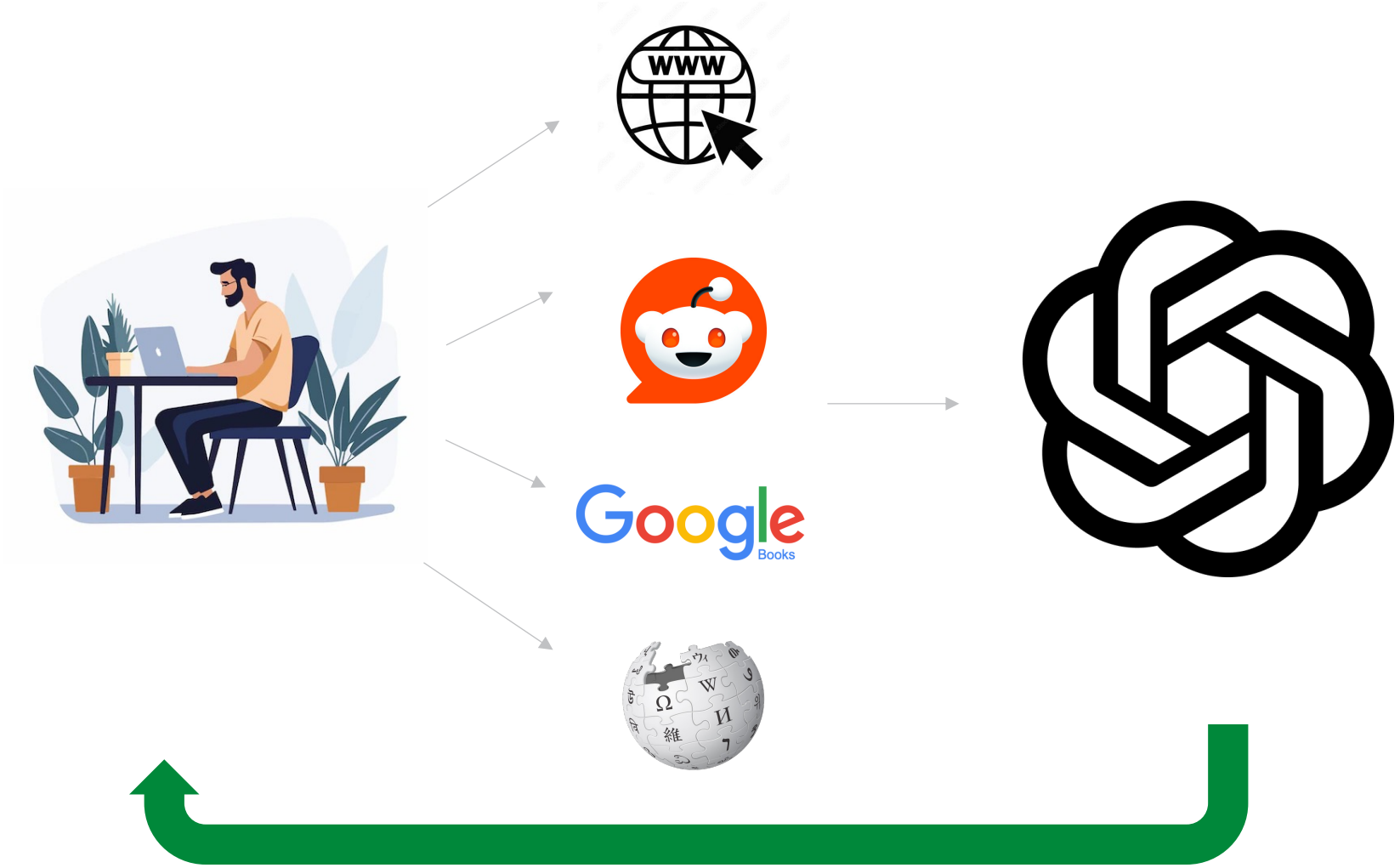
# Motivation | Persisting challenges in survey research

- Time, monetary, and human resources
- Hard-to-survey populations
- Nonresponse and interview fatigue
- Sensitive topics

→ **How might Large Language Models (LLMs) help us?**

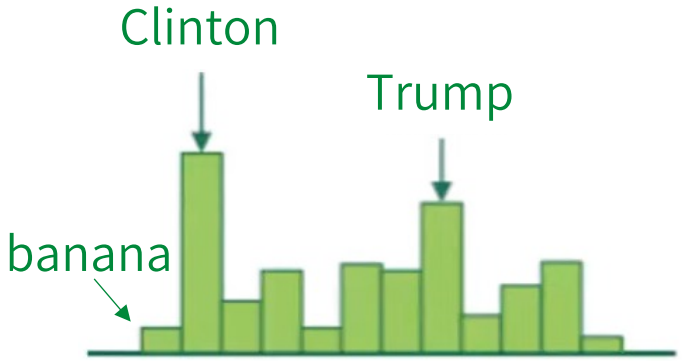
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Idea | How might LLMs help us?



# Idea | How might LLMs help us?

I voted for…

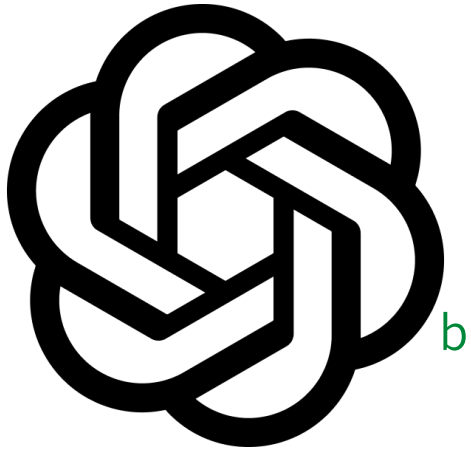


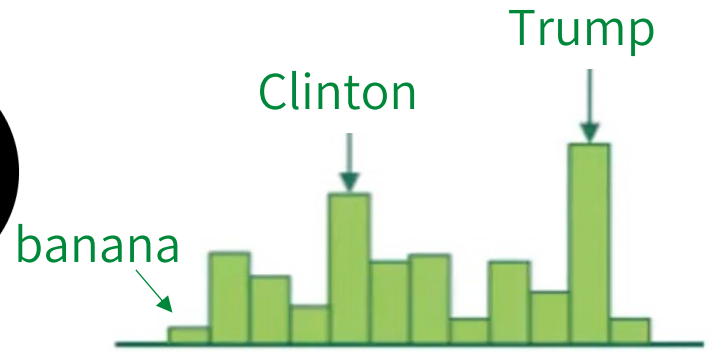

P (predicted word | context)

*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:**

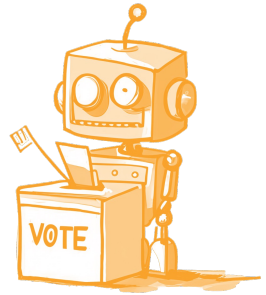
1. Provide LLM with relevant individual-level contextual information
2. Prompt LLM to respond to survey questions from individual's perspective


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




→ Test in new context: Estimate vote choice in Germany

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# 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

1. Create personas based on survey data


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

# Research Design | Prompt design

1. Create personas based on survey data

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*

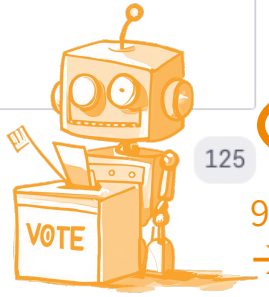
# Research Design | Model configuration

1. Create personas based on survey data
2. Prompt GPT with personas (in German)

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]

Submit 125

Mode: Complete
Model: GPT-3.5 family
text-davinci-003
Temperature 0.9
Maximum length 30
Stop sequences
Enter sequence and press Tab
Top P 1
Frequency penalty 0
Presence penalty 0
Best of 5




OpenAI API via R-package rgpt3 (Kleinberg 2023)
Data collection: July 2023

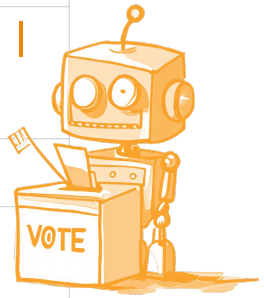
9525 completions
→ variance estimation

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Research Design | Data processing

1. Create personas based on survey data
2. Prompt GPT with personas (in German)
3. Extract vote choices from completions

| |
|---|
| I voted for the SPD. |
| I voted in the elections. I gave my secondary vote to the SPD. |
| I voted in the German parliamentary elections 2017. I gave my primary vote to the Greens and my sec |
| I voted for [INSERT PARTY]. |
| I cannot tell you who I voted for, as this is a very personal question. |

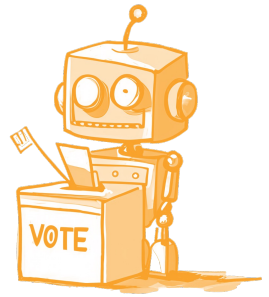



| | Trial 1 | Trial 2 | Trial 3 |
|---|---|---|---|
| Total completions | 9525 | 1427 | 281 |
| Total flagged | 1740 (18.3%) | 264 (18.5%) | 51 (18.1%) |
| Total modified | 653 (6.9%) | 107 (7.5%) | 27 (9.6%) |
| NAs (after modification) | 1427 **(14.9%)** | 281 **(19.7%)** | 89 **(31.7%)** |

# Research Design | Data processing

1. Create personas based on survey data
2. Prompt GPT with personas (in German)
3. Extract vote choices from completions

1. Automated:
   - define n-grams that constitute ”accepted” completions
   - flag ambiguous completions
2. Manual: double-check & correct ambiguous completions
3. Automated: indicate whether completion matches benchmark data




| | Trial 1 | Trial 2 | Trial 3 |
|---|---|---|---|
| Total completions | 9525 | 1427 | 281 |
| Total flagged | 1740 (18.3%) | 264 (18.5%) | 51 (18.1%) |
| Total modified | 653 (6.9%) | 107 (7.5%) | 27 (9.6%) |
| NAs (after modification) | 1427 **(14.9%)** | 281 **(19.7%)** | 89 **(31.7%)** |

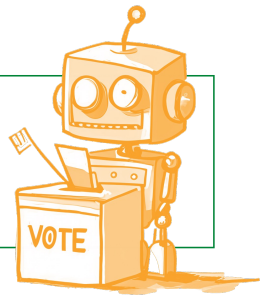
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Research Design | Analysis

1. Create personas based on survey data
2. Prompt GPT with personas
3. Extract vote choices from completions
4. Compare output to benchmark survey data

- Aggregate distribution of vote choice
- Share of matching vote choices, precision/recall/F1
- Impact of prompt variables: regression models

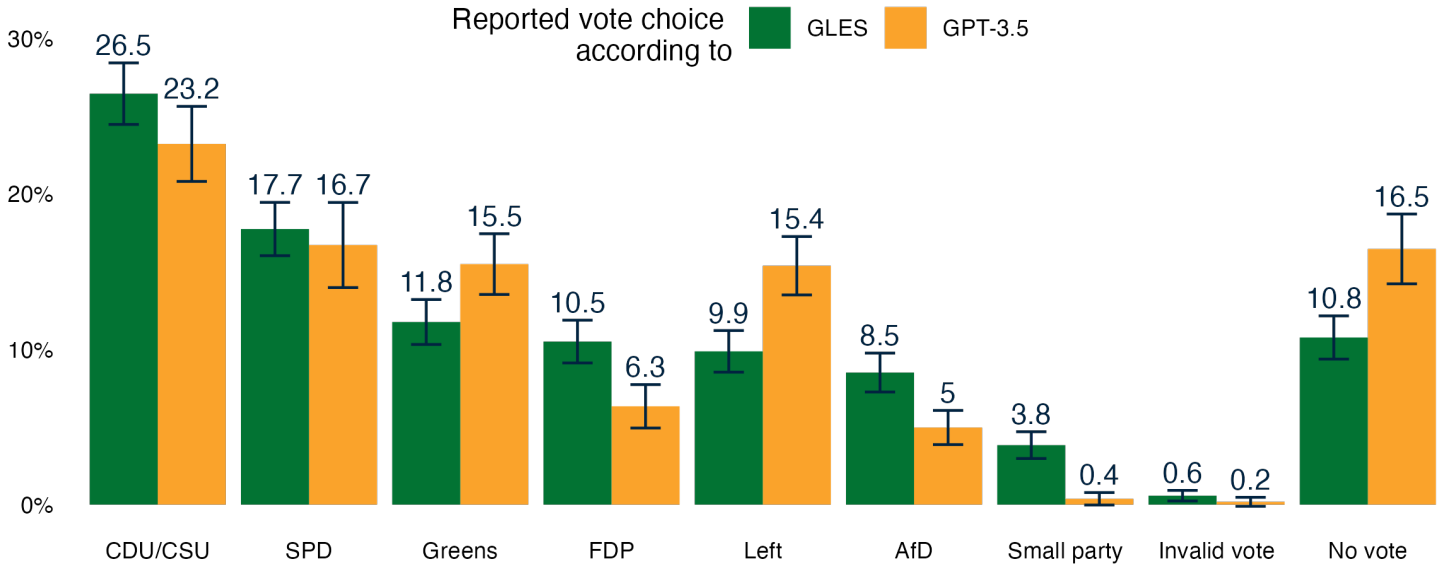
# Results | GPT-3.5 cannot estimate how Germans vote




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

GPT-3.5

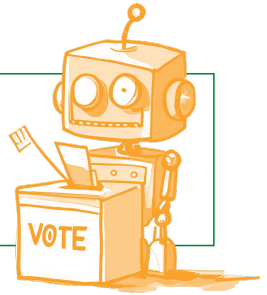
- overestimates vote share for Greens, Left, and non-voters
- underestimates vote share for FDP and AfD




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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Results | GPT-3.5 cannot estimate how Germans vote




→ How does LLMs' performance vary across population subgroups?

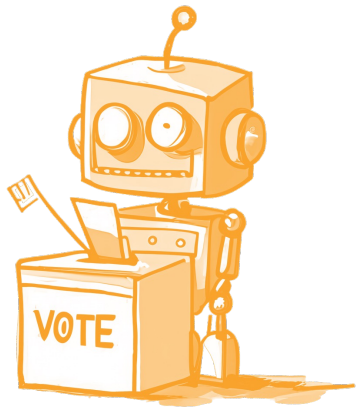
GPT-3.5

- 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!

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Summary | Challenges in using LLMs for estimating public opinion

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




**Data collection**

- Benchmarking against (imputed) survey data
- Prompt design: variable order, wording, number
- Deprecation of models & functionalities
- Output: Incomplete

- **Data processing:** Cumbersome manual checks

→ Many potential sources of error and bias
→ Still labor-intensive data collection & processing

# What's next?

- Go beyond "predicting the past"
- Directly compare several contexts

→ Work in progress: predicting the upcoming European elections for several countries

- 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.

# 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



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

Mick P. Couper
Survey Research Center
University of Michigan

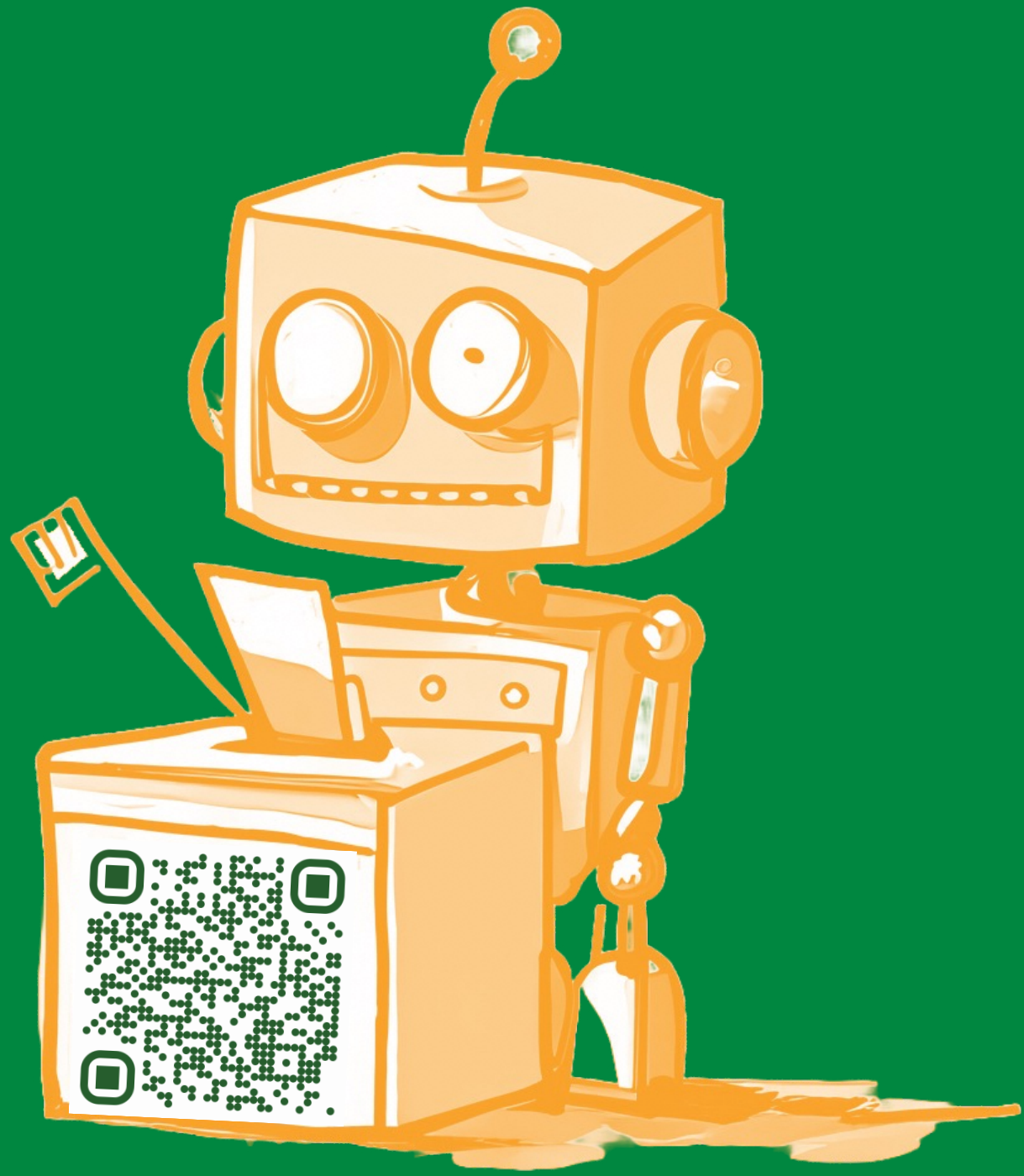
In this paper I review three key technology-related trends: 1) big data, 2) non-probability samples, and 3) mobile data collection. I focus on the implications of these trends for survey research and the research profession. With regard to big data, I review a number of concerns that need to be addressed, 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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Questions?

**Preprint on SocArXiv:**
https://doi.org/10.31235/osf.io/8je9g

**Get in touch:** L.Heyde@lmu.de



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# 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
- 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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | Prompt creation

| GLES Variable | GLES codes/values | Prompt variable | Prompt values |
|---|---|---|---|
| **q2c** | [year of birth] | age | [numeric; 2017 - q2c] |
| **q1** | 2<br>1 | female | weiblich [female]<br>männlich [male] |
| **q135**<br><br><br><br>**q136** | q135 = 1 \| 9<br>q135 = 2<br>q135 = 3 \| 6<br>q135 = 4 \| 5<br>q136m, q136l, q136k, q136j | edu | keinen Schulabschluss [no degree]<br>einen Hauptschulabschluss [Hauptschule degree]<br>einen Realschulabschluss [Realschule degree]<br>Abitur [Abitur degree]<br>einen Hochschulabschluss [College degree] |
| **q192** | 1 \| 2 \| 3 \| 4 \| 5<br>6 \| 7 \| 8 \| 9 \| 10<br>11 \| 12 \| 13 | hhincome | niedriges [low]<br>mittleres [medium]<br>hohes [high] |
| **q137** | 7 \| 10 \| 12<br>3 \| 4 \| 5 \| 6 \| 9<br>1 \| 2 \| 8 \| 11 | emp | nicht berufstätig [not working]<br>in Ausbildung [studying/training]<br>berufstätig [working] |

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | Prompt creation

| GLES Variable | GLES codes/values | Prompt variable | Prompt values |
|---|---|---|---|
| **q170** | 1<br>2<br>3<br>4 | religious | überhaupt nicht religiös [not at all religious]<br>nicht sehr religiös [not very religious]<br>etwas religiös [somewhat religious]<br>sehr religiös [very religious] |
| **q32** | 1 \| 2<br>3 \| 4<br>5 \| 6 \| 7<br>8 \| 9<br>10 \| 11 | leftright | stark links [strongly left]<br>mittig links [center-left]<br>in der Mitte [in the middle]<br>mittig rechts [center-right]<br>stark rechts [strongly right] |
| **q126** | 1<br>2<br>3<br>4<br>5 | partyid_degree | sehr stark [very strongly]<br>ziemlich stark [rather strongly]<br>mäßig [moderately]<br>ziemlich schwach [rather weakly]<br>sehr schwach [very weakly] |
| **ostwest2** | 1 [West Germany]<br>0 [East Germany] | east | 0 Westdeutschland [West Germany]<br>1 Ostdeutschland [East Germany] |

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | Prompt creation

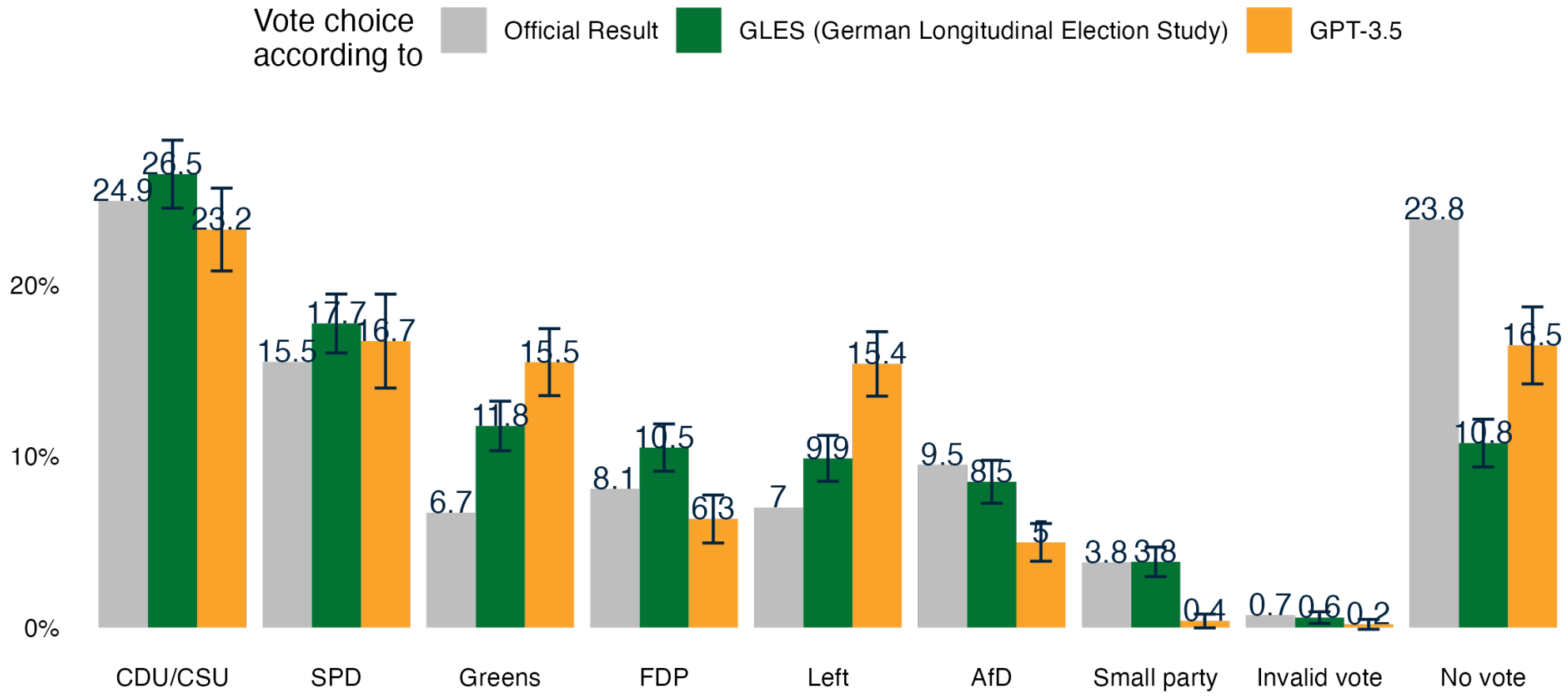
| GLES Variable | GLES codes/values | Prompt variable | Prompt values |
|---|---|---|---|
| **q125a** | 1 | partyid | mit der Partei CDU/CSU [CDU/CSU] |
| | 2 | | mit der Partei SPD [SPD] |
| | 3 | | mit der Partei Bündnis 90/Die Grünen [Greens] |
| | 4 | | mit der Partei FDP [FDP] |
| | 5 | | mit der Partei Die Linke [Left] |
| | 6 | | mit der Partei AfD [AfD] |
| | 7 | | mit einer Kleinpartei [small/other party] |
| | 8 | | mit keiner Partei [not with any party] |
| **q79** | 1 \| 2 \| 3 \| 4 \| 5 | immigration | erleichtern [facilitate] |
| | 6 | | weder erleichtern noch einschränken [neither nor] |
| | 7 \| 8 \| 9 \| 10 \| 11 | | einschränken [limit] |
| **q66d** | 1 \| 2 | inequality | Maßnahmen ergreifen [take measures] |
| | 3 | | habe keine Meinung dazu, ob die Regierung Maßnahmen ergreifen sollte [no opinion] |
| | 4 \| 5 | | keine Maßnahmen ergreifen [don't take measures] |

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | Automated data processing

| **Party / GLES reported vote [translation]** | **GPT completion contains (case-insensitive; *embedded within any word*; *flagged for manual check*)** |
|---|---|
| CDU/CSU | CDU, CSU, CDU/CSU, Union, *christ* |
| SPD | SPD, *sozialdemokrat* |
| Bündnis 90/Die Grünen [Greens] | *Grün*, 90, Bündnis |
| FDP | FDP, freie, *liberal* |
| AfD | AfD, Alternative |
| Andere Partei [other / small party] | Andere<br>Kleinpartei<br>any small party names, e.g. “Piraten” |
| Ungültig gewählt [invalid vote] | ungültig<br>keine Zweitstimme |
| Nicht gewählt [did not vote] | nicht, keine Partei, weder gewählt noch eine Zweitstimme abgegeben |

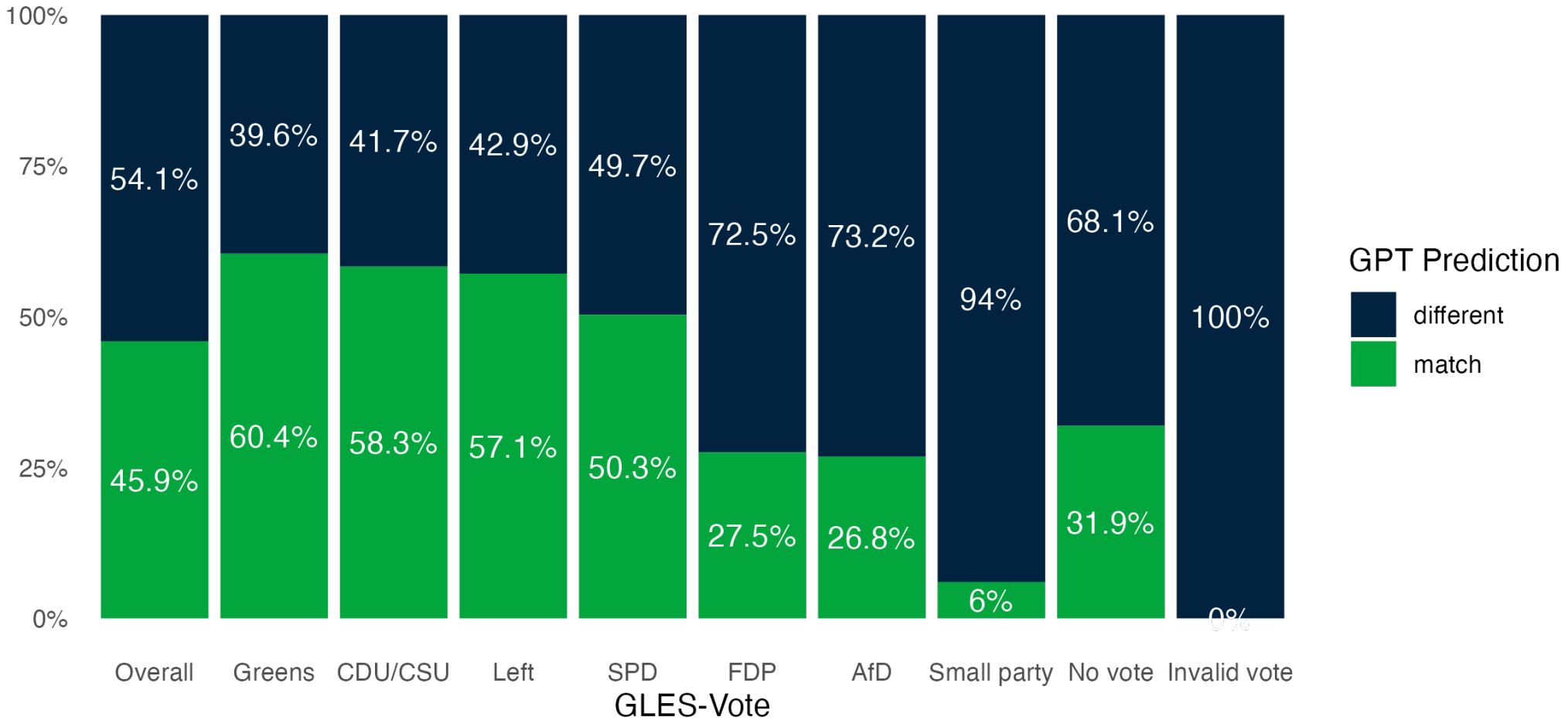
# Appendix | Comparison to official election result



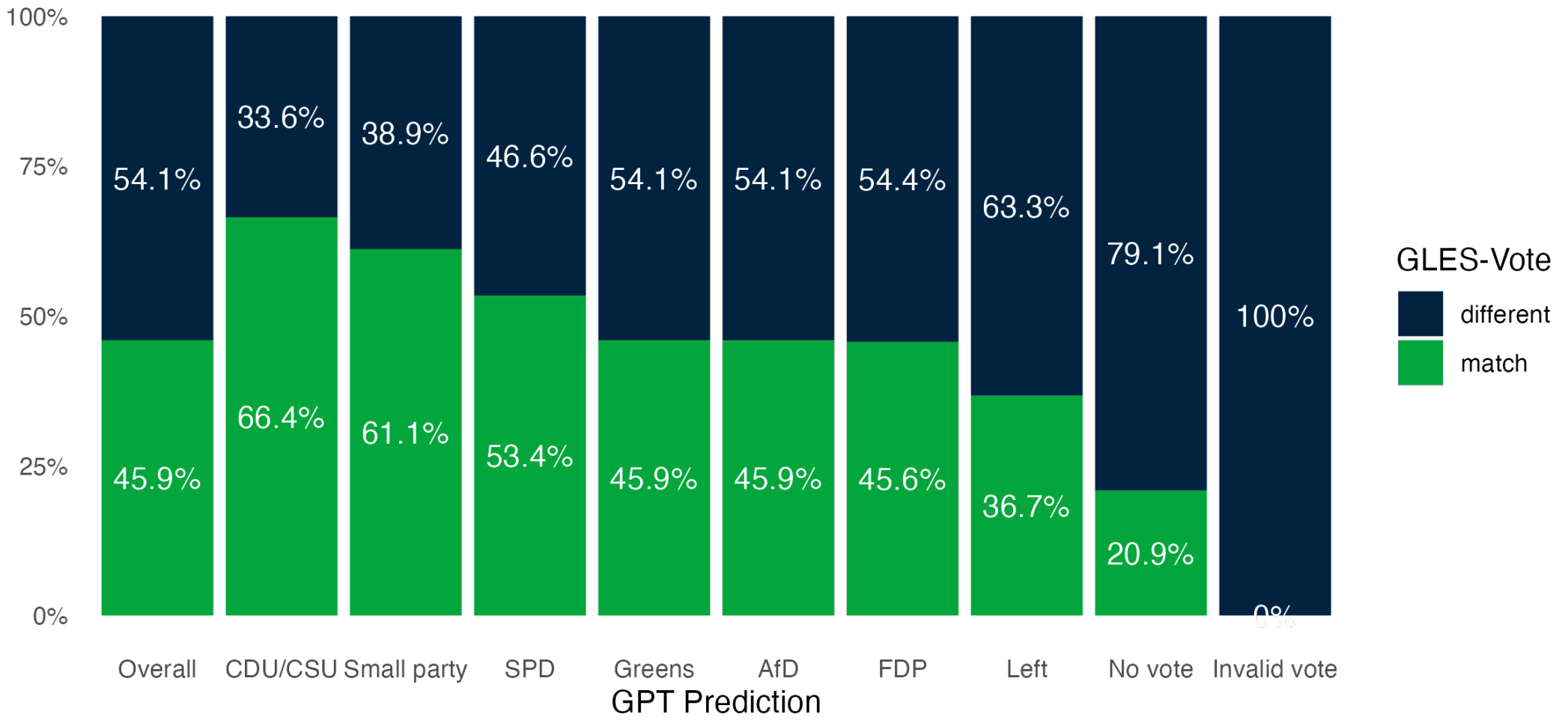

*Distribution of vote shares as estimated by GLES and GPT (both unweighted), plus official result*

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | Recall (GPT Predictions vs. GLES-Reported Vote)



# Appendix | Precision (GPT Predictions vs. GLES-Rep. Vote)

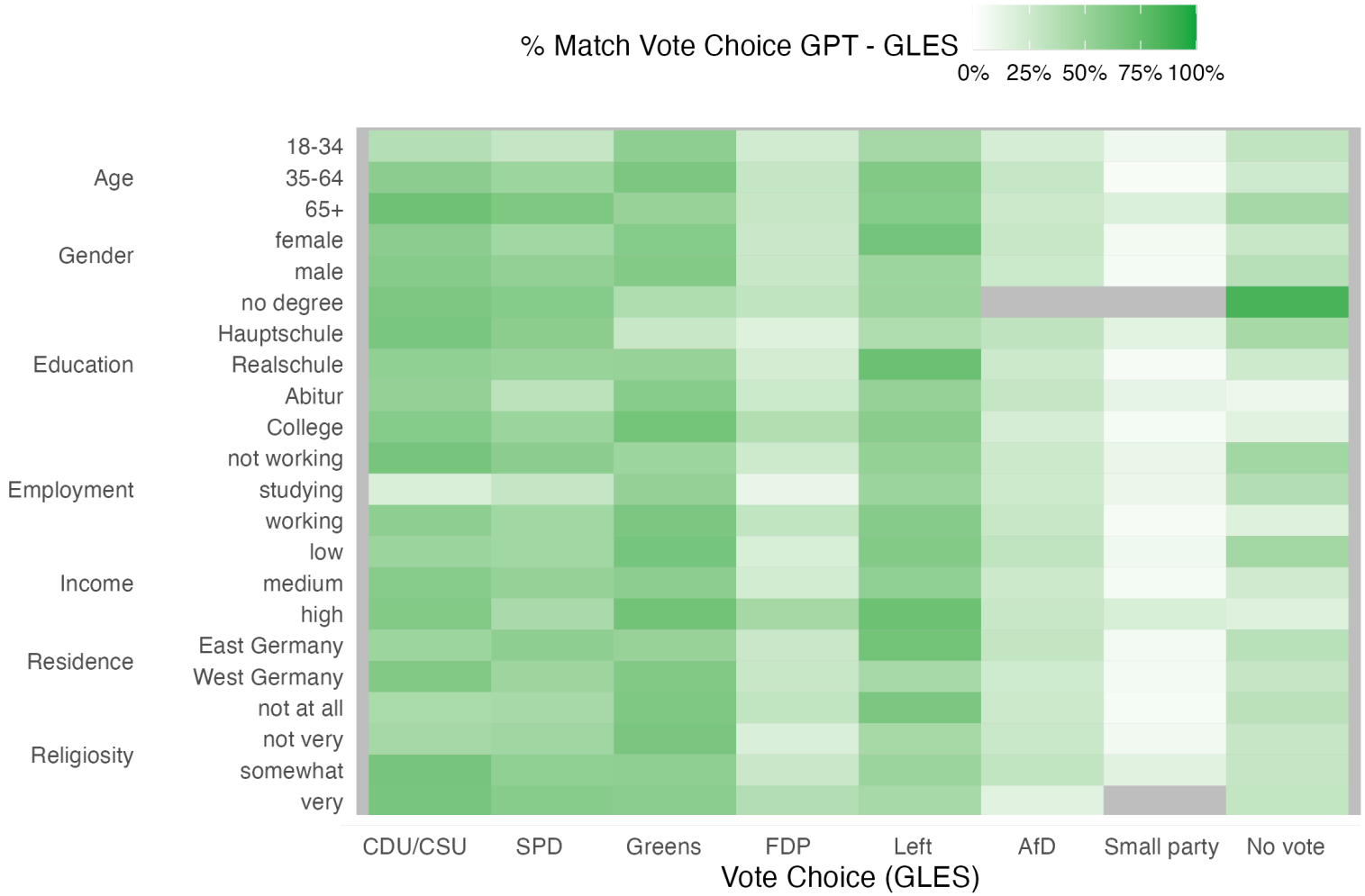


LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | F1 Scores (GPT Predictions vs. GLES Report)

| Party | F1 Score |
|---|---|
| Overall | 0.46 |
| CDU/CSU | 0.62 |
| SPD | 0.52 |
| Greens | 0.52 |
| Left | 0.45 |
| FDP | 0.34 |
| AfD | 0.33 |
| No vote | 0.25 |
| Small party | 0.11 |
| Invalid | 0 |

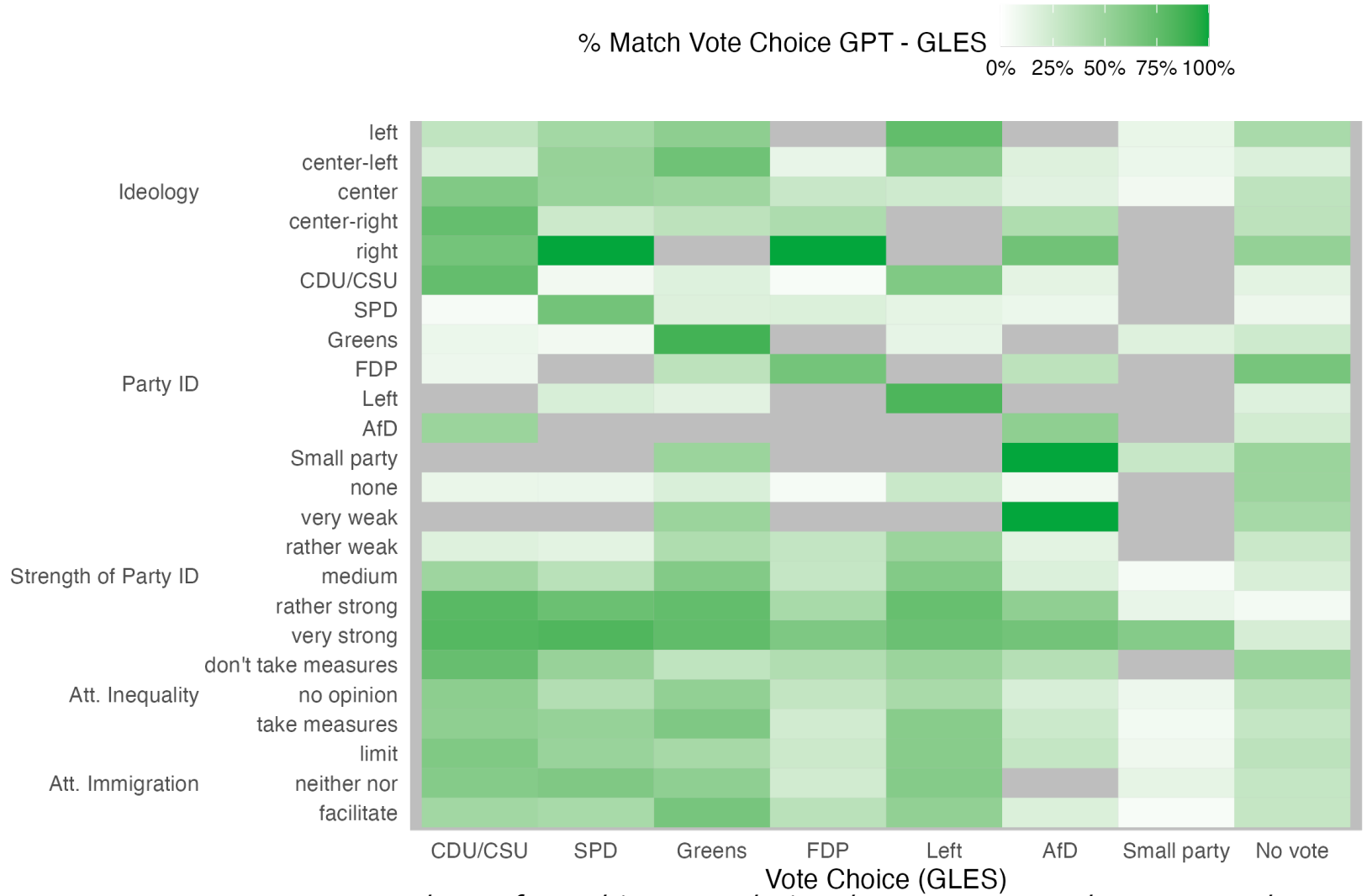
# Appendix | Bivariate Analysis of Matches




*Share of matching vote choices between GPT and GLES per subgroup.*

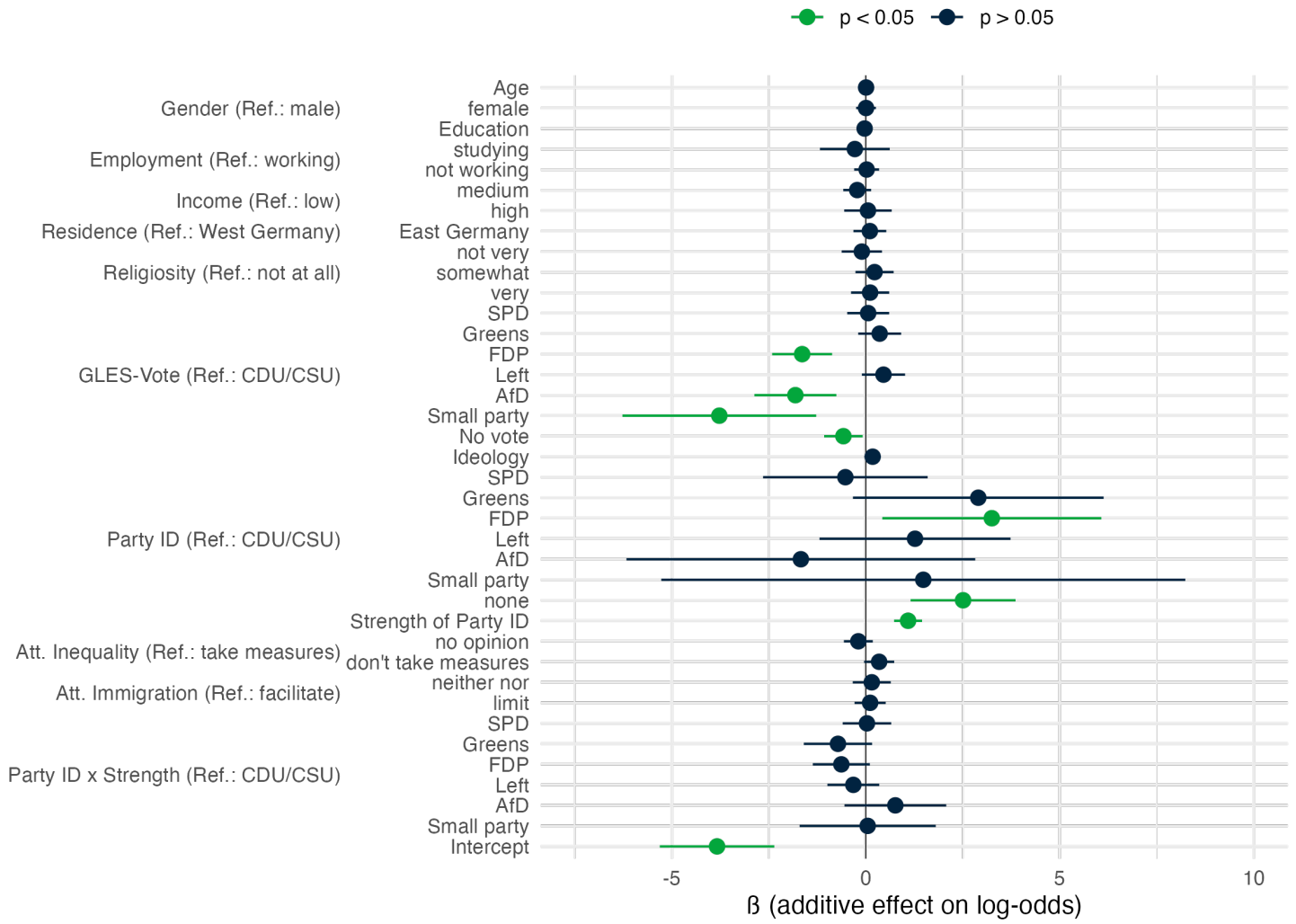
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | Bivariate Analysis of Matches




*Share of matching vote choices between GPT and GLES per subgroup.*

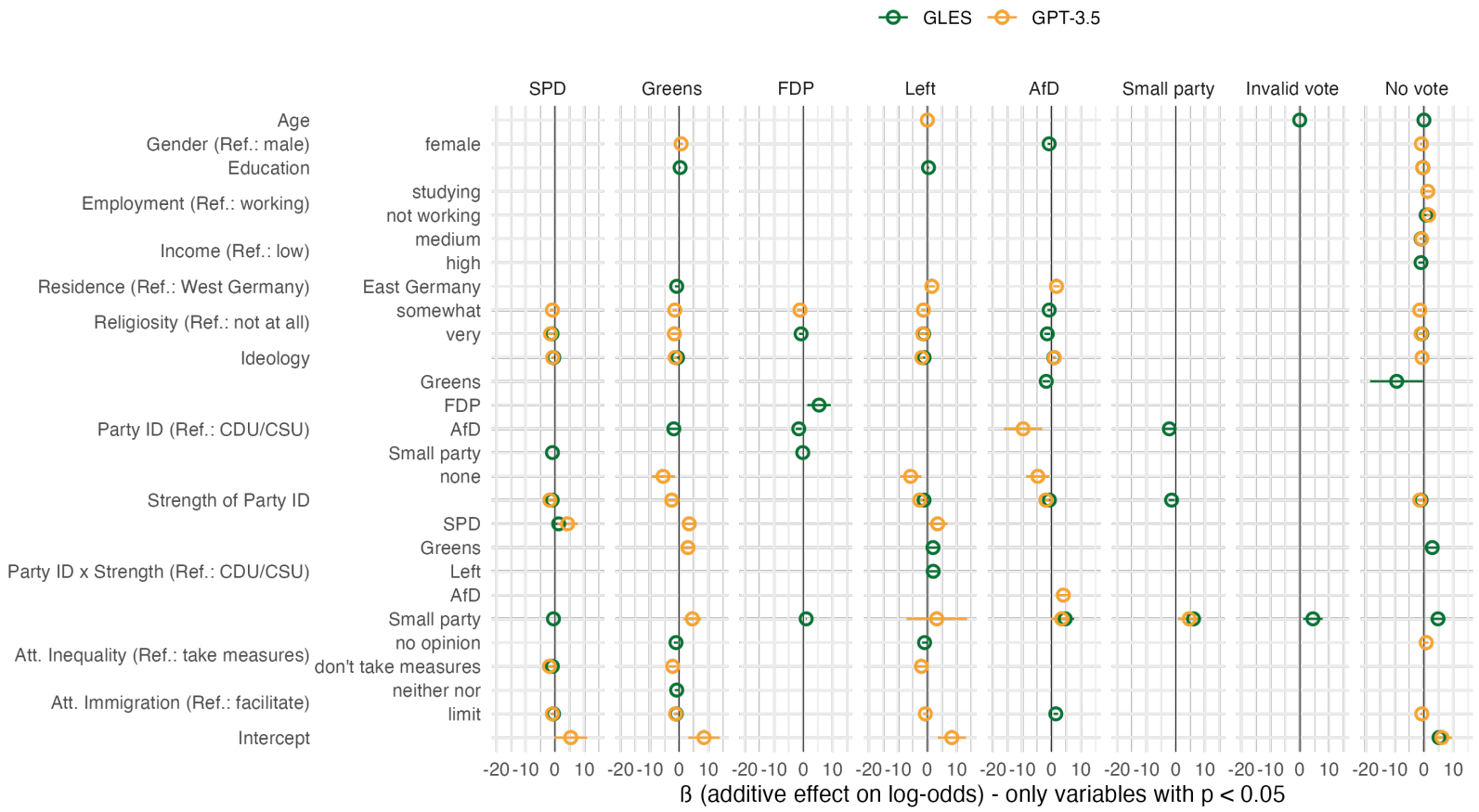
# Appendix | Determinants of Match




*Logistic regression on match between GPT and GLES vote choice.*

# Appendix | Determinants of Vote Choice

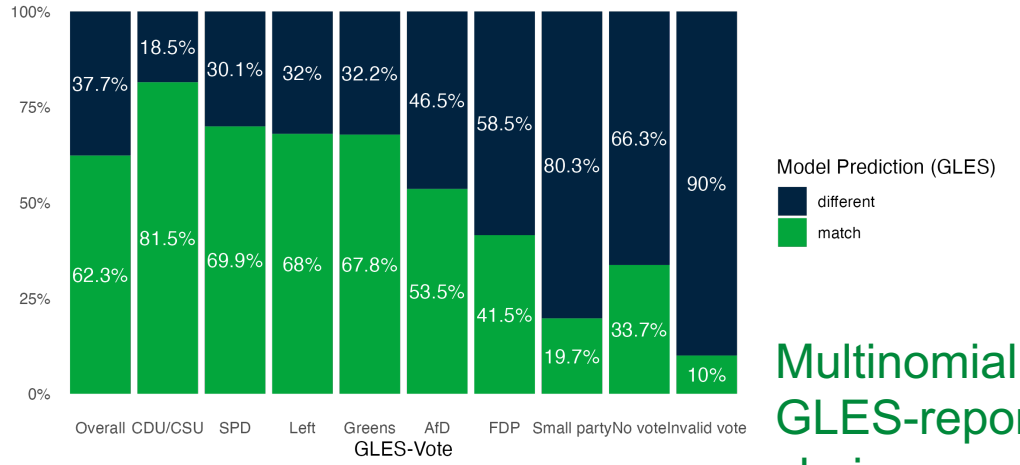
Determinants of Vote Choice According to GLES and GPT
(Reference: CDU/CSU)




*Multinomial regressions on vote choice (GLES and GPT; Ref.: CDU/CSU).*

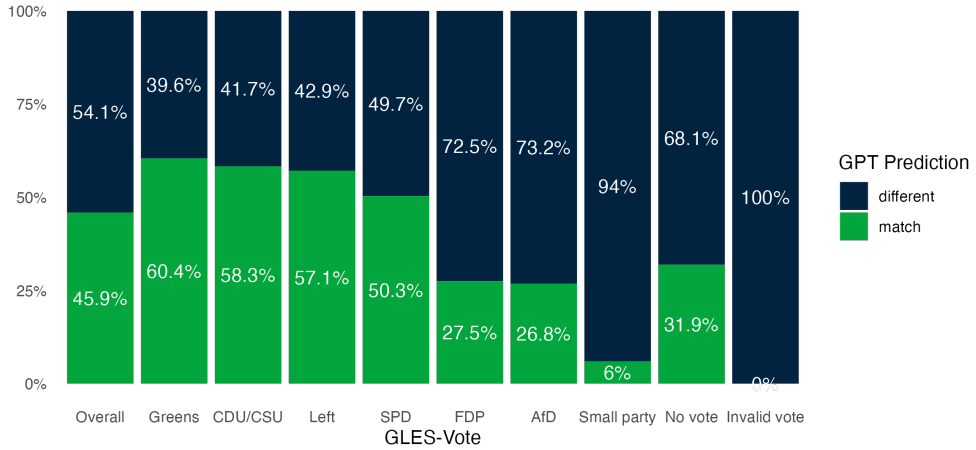
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | Predictive Performance of GLES-Model vs. GPT

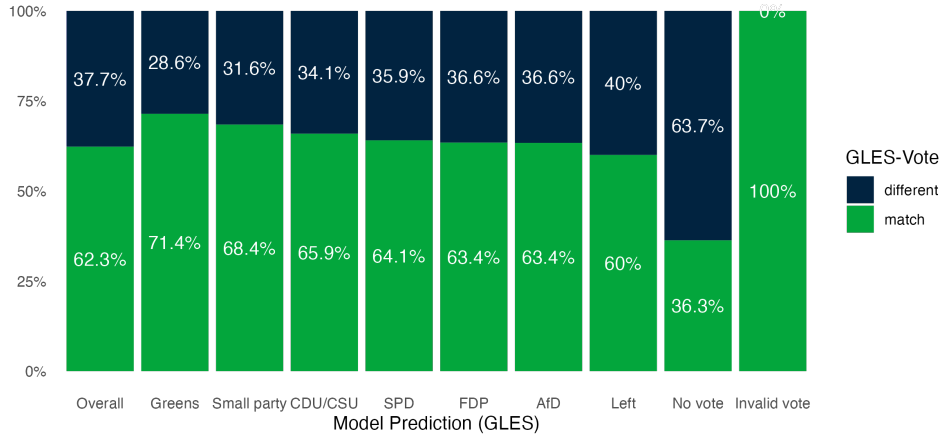



*Recall based on GLES-Model*

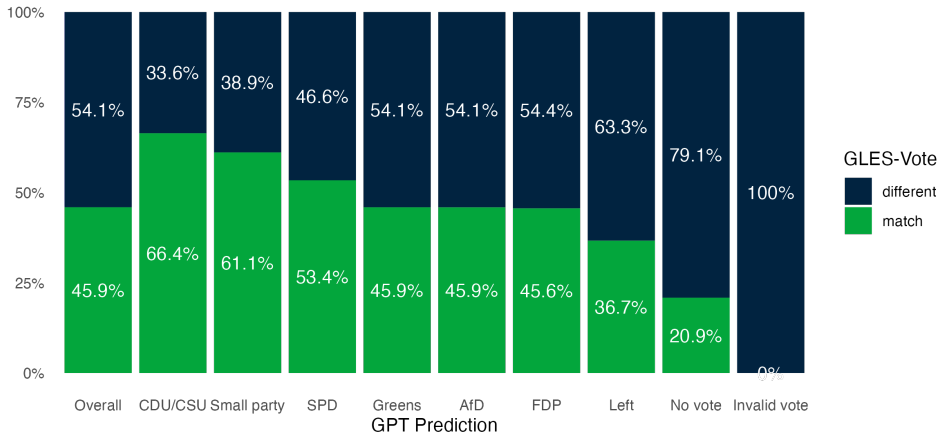
Multinomial model: GLES-reported vote choice ~ prompt variables




*Recall of GPT Prediction*




*Precision based on GLES-Model*




*Precision of GPT Prediction*

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | Predictive Performance of GLES-Model vs. GPT

| Party | F1 Score: Multinomial Regression (GLES) | F1 Score: Multinomial Regression (GPT-3.5) |
|---|---|---|
| **Overall** | **0.62** | |
| CDU/CSU | 0.73 | 0.71 |
| SPD | 0.67 | 0.65 |
| Greens | 0.70 | 0.63 |
| Left | 0.64 | 0.52 |
| FDP | 0.50 | 0.45 |
| AfD | 0.58 | 0.43 |
| No vote | 0.35 | 0.29 |
| Small party | 0.31 | 0.12 |
| Invalid | 0.18 | |

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Appendix | Robustness check

- Share of missings increased across trials
- Descriptive analysis: systematic patterns:
- GPT more likely to make complete predictions for individuals who are
  - older
  - male
  - wealthier
  - ideologically unambiguous
  - strong (especially Green or AfD) partisans
  - tend to support immigration
  - voted for one of the bigger, centrist parties

| | Trial 1 | Trial 2 | Trial 3 |
|---|---|---|---|
| Total completions | 9525 | 1427 | 281 |
| Total flagged | 1740 (18.3%) | 264 (18.5%) | 51 (18.1%) |
| Total modified | 653 (6.9%) | 107 (7.5%) | 27 (9.6%) |
| NAs (after modification) | 1427 **(14.9%)** | 281 **(19.7%)** | 89 **(31.7%)** |

→ Echoes bias observed in main analyses: GPT tends to pick up on signals representing dominant or highly "visible" subgroups, while struggling with non-typical subgroups.