LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Can Large Language Models Predict How People Vote?
Evidence from Germany

Anna-Carolina Haensch
LMU Munich
University of Maryland, College Park

work with
Leah von der Heyde (LMU Munich) and
Alexander Wenz (University of Mannheim)


How can I deal with missing data?


How can I create a synthesized version of my data?


What can we do about interview fatigue and declining response rates?

These questions are not new!

Questions about missing data, sensitive information and data synthesis, declining response rates have been around for a long time in (social) data science and survey statistics.

What is indeed new is that they can all now be tackled with Generative AI and Large Language Models.

But should they?

Disclaimer

This is not a talk about all useful use cases of LLMs in survey statistics, but we focus mainly on tasks which involve predicting, synthesizing or imputing variable values of members in a (human) population.


We can now use LLMs to impute or synthesize values from surveys and other data collection types!
This a ridiculous idea, LLMs were never built for such tasks!
Let‘s find out!

What are quality criteria for synthetic* values?

Statistical similarity

Feature/Variable correlations

Fidelity in general

(Privacy-preserving metrics)

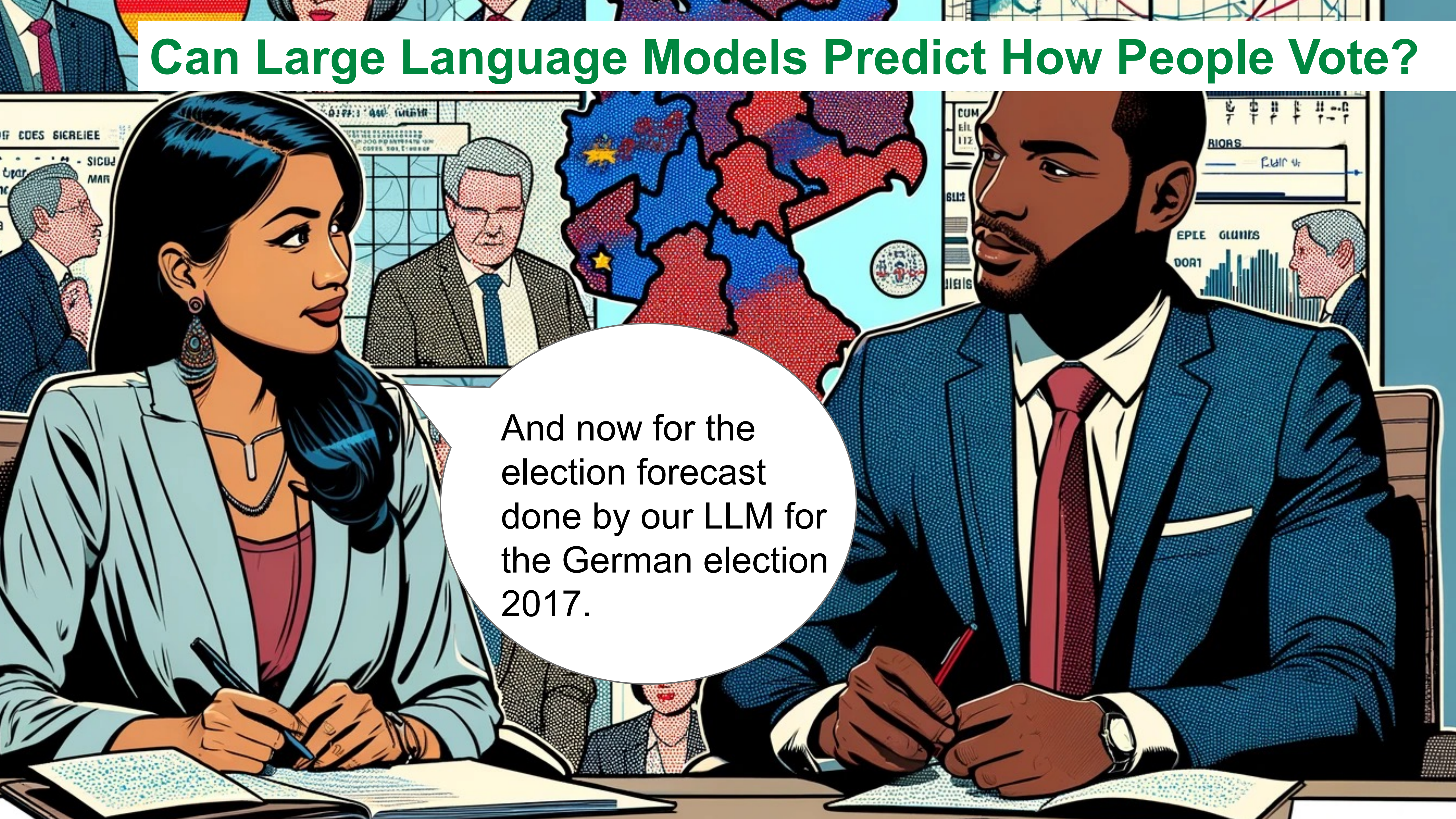
*We could also focus on prediction or imputation tasks and their quality indicators of course, but for now we will focus on data synthesis.

In many cases, it is helpful to set an external data set as ground truth.

→ **This is what we will do in our exploratory study**

→ Let‘s move on to our concrete example!

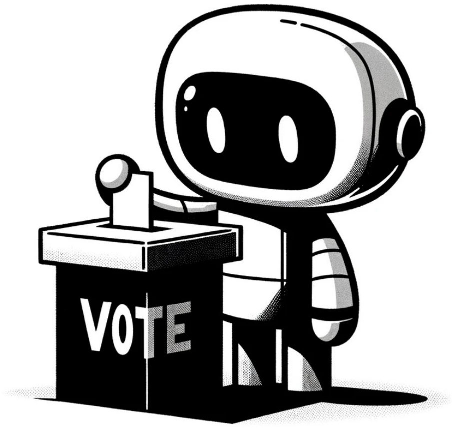
Can Large Language Models Predict How People Vote?



Our test case:

Can GPT-3.5 accurately synthesize individual voting behavior?

How does its result compare with actual survey data from the German Longitudinal Election Study (GLES)?



The comparison data:

- Dataset: 2017 post-election cross-section of the GLES
- Sample: Voting-eligible participants reporting their vote choice
- Key Variables:
  - Demographics: Age, gender, educational attainment, income, employment status, residence in East/West Germany
  - Political Views: Religiosity, ideological left-right self-placement, strength of political partisanship, attitudes towards immigration and income inequality
- Imputation: Missing values imputed for 20% of respondents

An example prompt - German (as prompted)

Ich bin 28 Jahre alt und weiblich. Ich habe einen Hochschulabschluss, ein mittleres monatliches Haushalts-Nettoeinkommen und bin berufstätig. Ich bin nicht religiös. Politisch-ideologisch ordne ich mich mittig links ein. Ich identifiziere mich ziemlich schwach mit der Partei Bündnis 90/Die Grünen. Ich lebe in Westdeutschland. Ich finde, die Regierung sollte die Einwanderung erleichtern und Maßnahmen ergreifen, um die Einkommens-unterschiede zu verringern. Habe ich bei der Bundestagswahl 2017 gewählt und wenn ja, welcher Partei habe ich meine Zweitstimme gegeben? Ich habe [INSERT]

An example prompt - 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]

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

8 more variables in the Appendix

GLES

German Longitudinal Election Study

13 variable values for 1,905 voting-eligible participants in the 2017 post-election cross-section of the GLES

I am xx years old … which party did I vote for? I [INSERT]

Created 1,905 prompts by inserting the values into our prompt template and prompt GPT-3.5 through the OpenAI API

I voted for the party xx.

Gave back 1,905 (x5) responses to the authors‘ prompts

The LLM

- API Usage: Employed **OpenAI's GPT-3.5 API, text-davinci-003 version**
- Temperature Setting: Calibrated to 0.9
- Token Limit: Responses capped at 30 tokens
- Data Collection Date: July 2023

Exemplary output from the OpenAI API

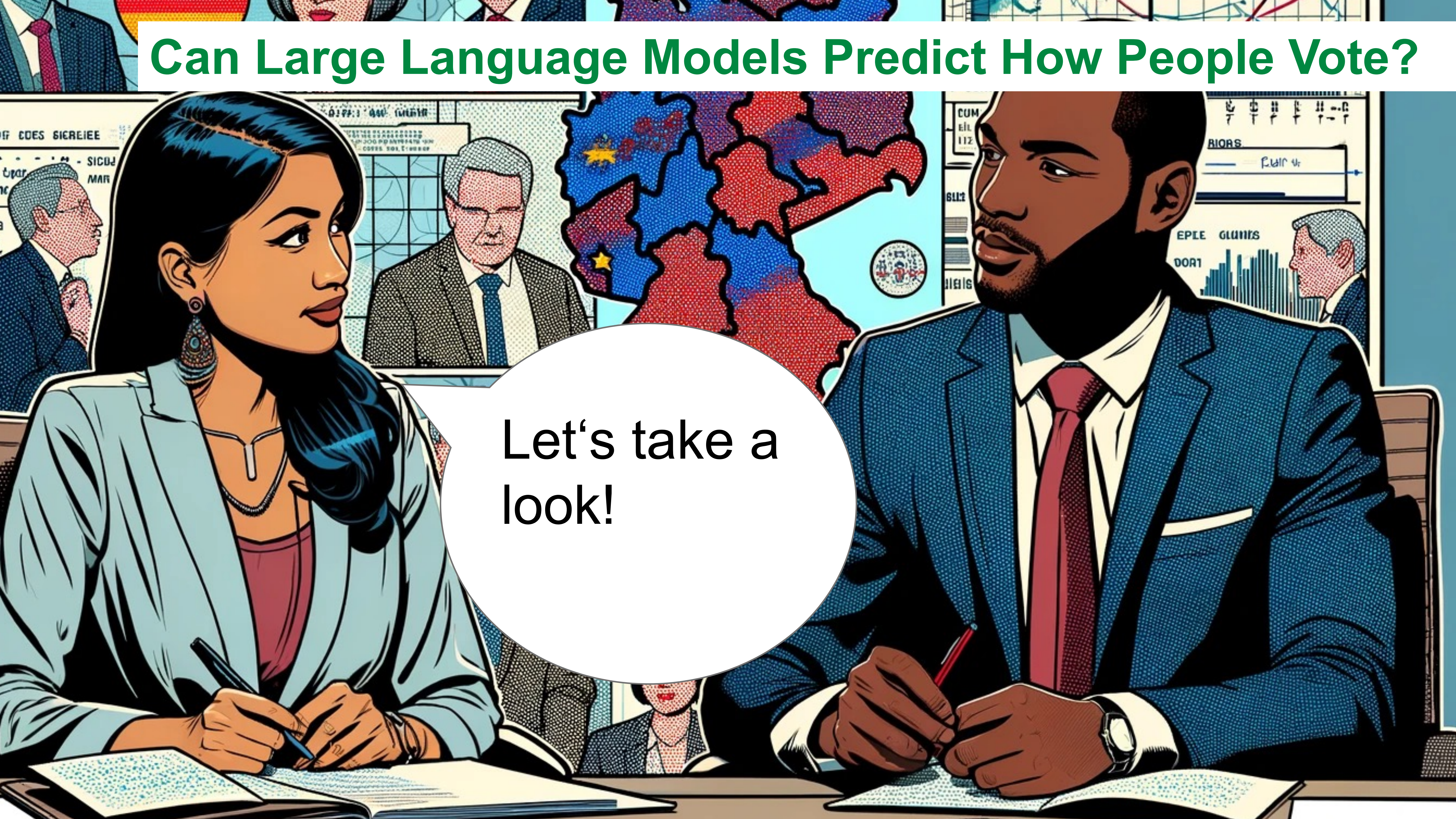
Ich habe für die SPD gewählt.
Ich habe der SPD meine Zweitstimme gegeben.
Ich habe Die Linke gewählt.
bei der Bundestagswahl 2017 gewählt und meine Zweitstimme der Partei Die Linke gege

I voted for the party xx.

Gave back 1,905 (x5) responses to the authors‘ prompts

Party / GLES reported vote (translation in brackets)	**GPT completion contains (case-insensitive; *asterisk*: embedded within any word)**
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 [confirmed by manual check]
Andere Partei [other / small party]	Andere [confirmed by manual check] Kleinpartei [confirmed by manual check] any small party names, e.g. Piraten [confirmed by manual check]
Ungültig gewählt [invalid vote]	[confirmed by manual check] ungültig keine Zweitstimme
Nicht gewählt [did not vote]	[confirmed by manual check] nicht, keine Partei, weder gewählt noch eine Zweitstimme abgegeben

Can Large Language Models Predict How People Vote?



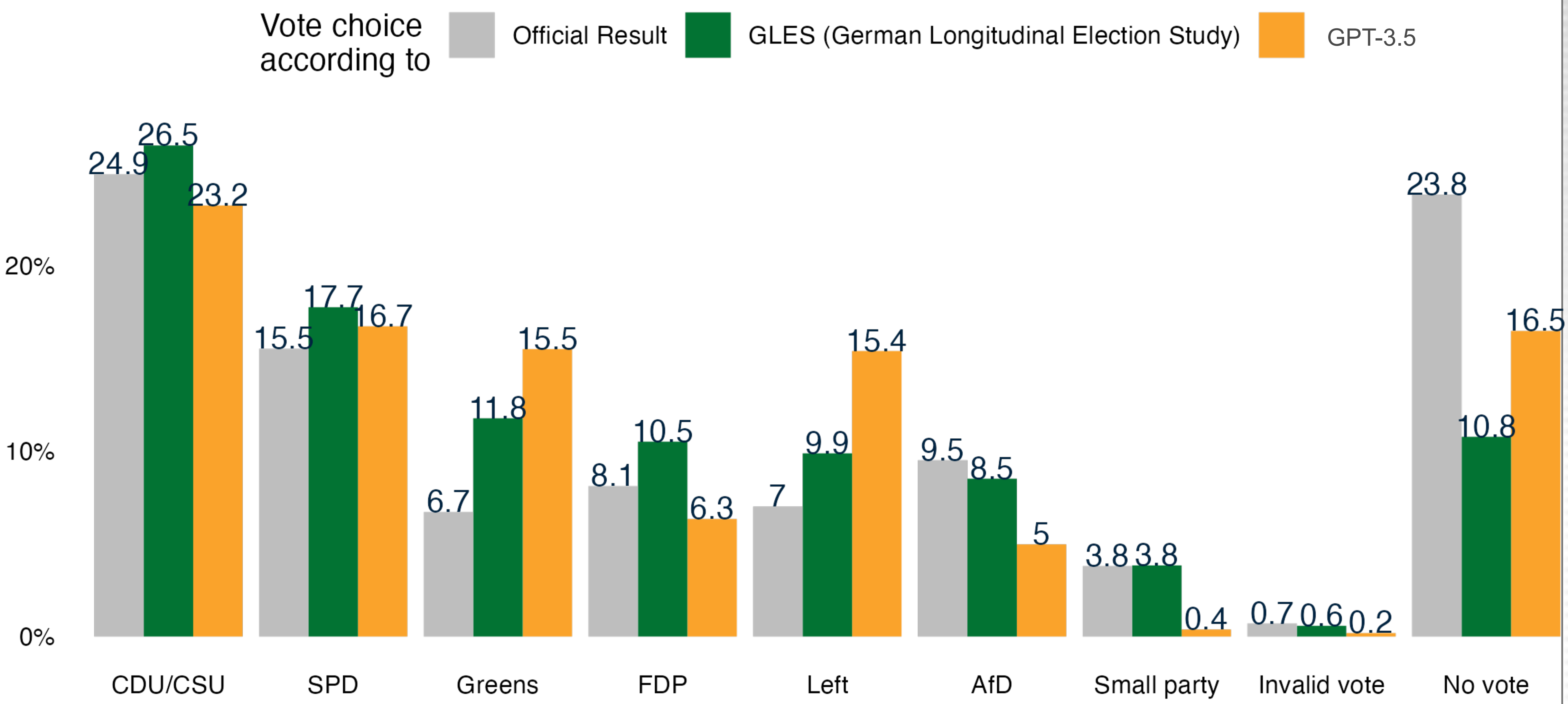


Figure 1: Distribution of vote shares as estimated by GLES and GPT(unweighted)

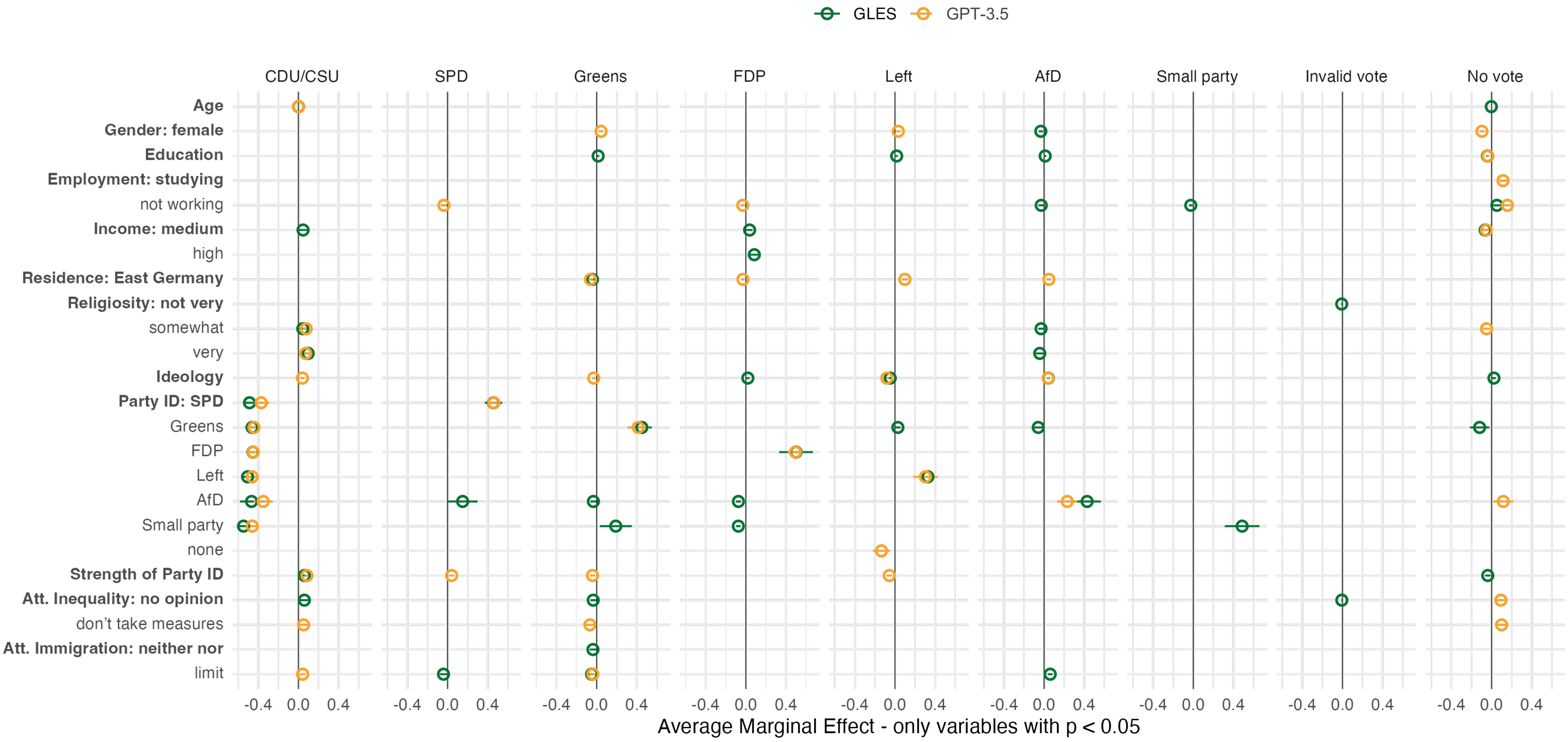


Figure 3: Multinomial regressions on vote choice, for GLES and GPT. Average Marginal Effects on Vote Choice.


We can now use LLMs to impute or synthesize values from surveys and other data collection types!
This a ridiculous idea, LLMs were never built for such tasks!
So, all in all, where do we stand?

Results

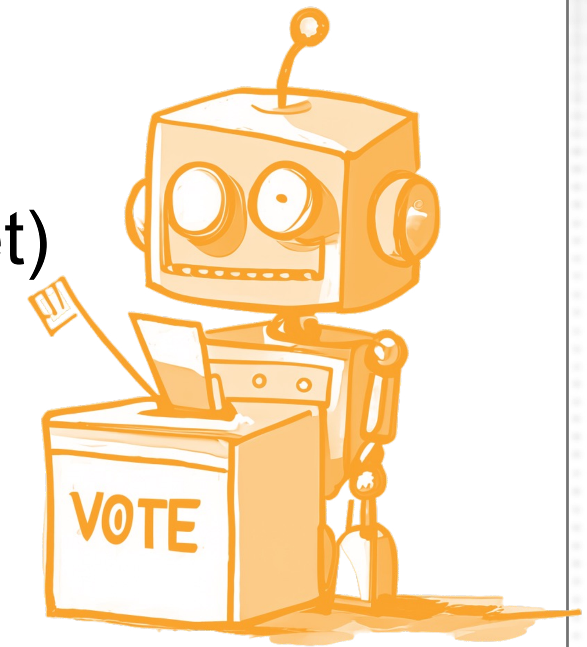
Comparative Analysis: GPT overestimated votes for Greens, Left, and non-voters, but underestimated FDP and AfD votes. Overall prediction accuracy was modest at 0.46.

Predictive Discrepancies: GPT outputs are correlated as expected with straightforward indicators like party identification or ideology but faltered on complex variables such as attitudes towards immigration, economic policy, and socio-demographics.

Limitations in Subgroup Prediction: Struggled with predicting votes for right-wing AfD and voters outside "typical" partisan characteristics.

Limitations

- **Prompt design impact**
- **API functionalities limitation**, inability to record token completion probabilities
- **Usage of GPT-3.5** (replicability, fast-moving target)
- **Survey data as benchmark**

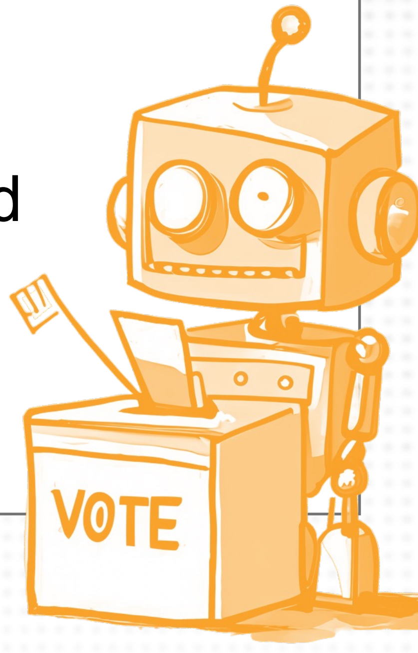


Summary

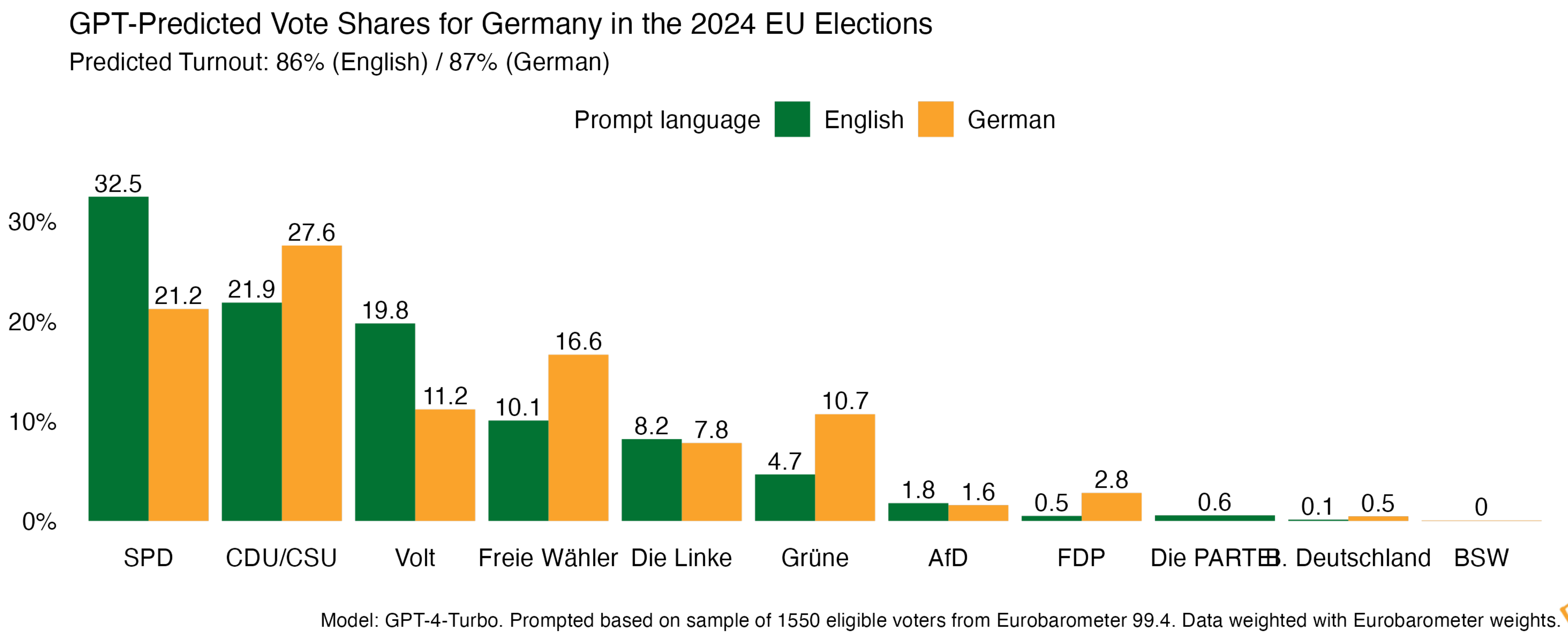
Estimation/prediction: GPT currently unsuitable for public opinion estimation.

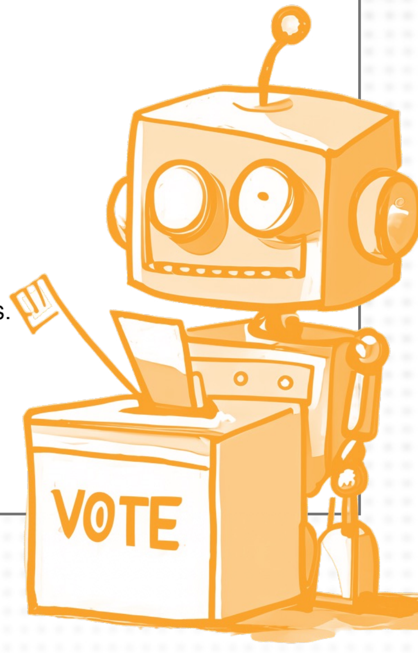
Synthesis: Depends on context but beware of

- Bias Against Subgroups: LLM fails to capture nuances among diverse voter groups, biased against certain subpopulations.
- Cross-National Performance: GPT's estimation less accurate in Germany compared to the U.S., likely poorer in less represented contexts.

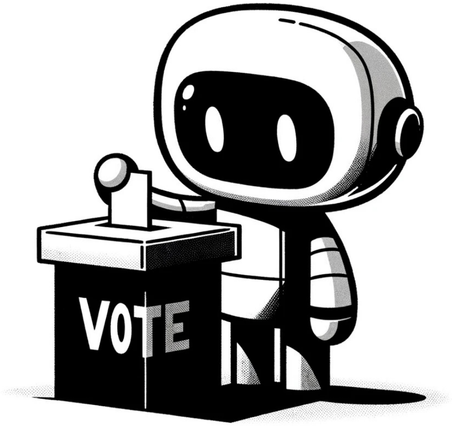


Outlook European Elections 2024





LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN



Thank you for your attention!

Grab it as a paper here:

Leah von der Heyde, Anna-Carolina Haensch, Alexander Wenz:
Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion.

C.Haensch@lmu.de