United in Diversity?
Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections

Leah von der Heyde
LMU Munich | Munich Center for Machine Learning | University of Mannheim

WAPOR Webinar | March 13, 2025

work with Anna-Carolina Haensch, Alexander Wenz, Bolei Ma

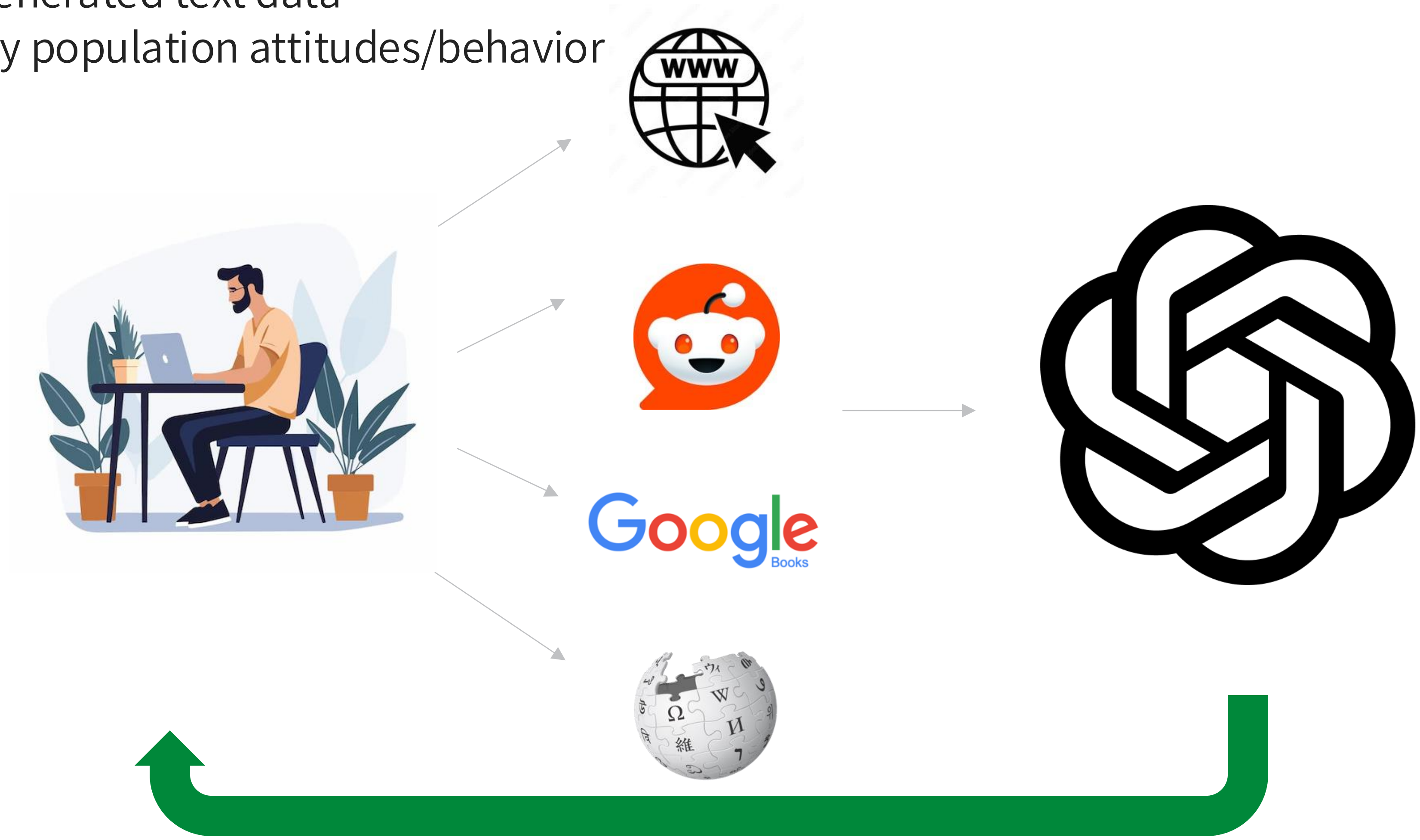
Motivation | Challenges in survey research

- **Time, monetary, and human resources** vs. predicting **future** outcomes short-notice
- Pre-testing & pilot studies
- Hard-to-survey populations
- Nonresponse and interview fatigue
- Sensitive topics

→ LLMs to the rescue?

Idea | Use characteristics of LLMs

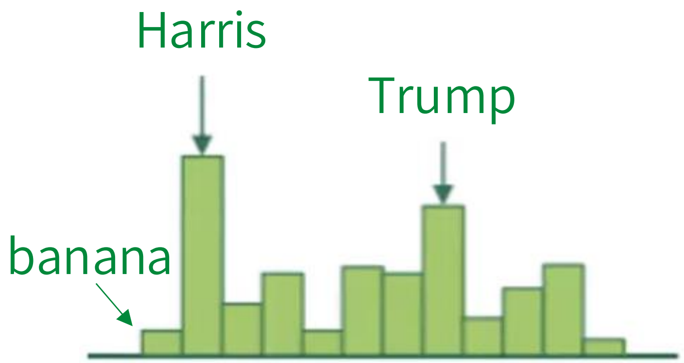
1. LLMs are trained on human-generated text data
   → potentially reflecting survey population attitudes/behavior



Idea | Use characteristics of LLMs

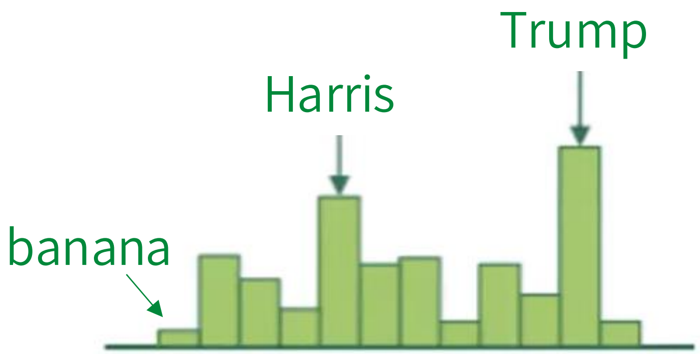
2. Output is conditional on training data AND prompt input

I voted for…



P (predicted word | context)

I am a Republican.
I voted for…



P (predicted word | context)

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

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Good Idea!? | Use LLMs to simulate survey respondents

Out of One, Many: Using Language Models to Simulate Human Samples

Lisa P. Argyle[1], Ethan C. Busby[1], Nancy Fulda[2], Joshua R. Gubler[1], Christopher Rytting[2] and David Wingate[2]

Language models trained on media diets can predict public opinion

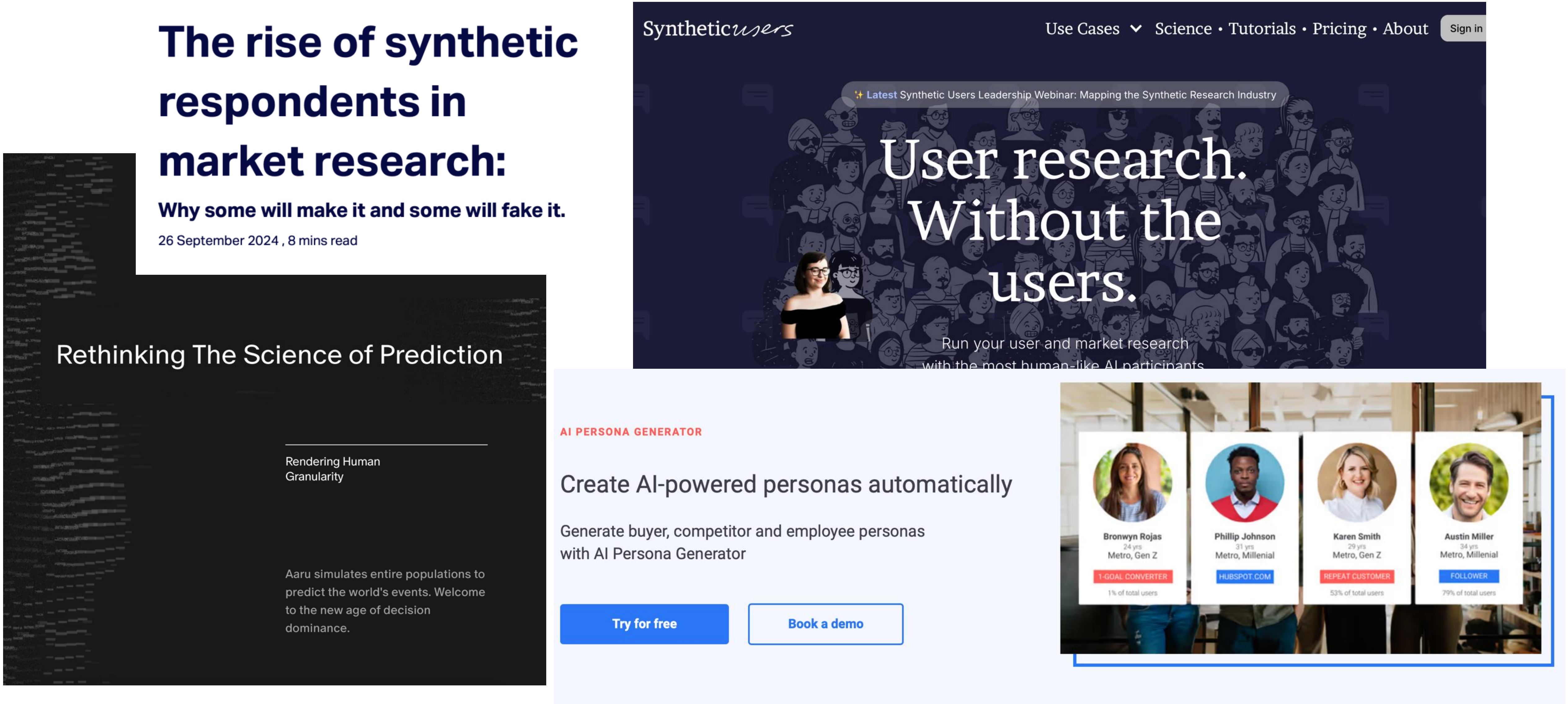
Eric Chu *†, Jacob Andreas[1], Stephen Ansolabehere[2], and Deb Roy[1]

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

Good Idea!? | Use LLMs to simulate survey respondents



Bad Idea!? | Use LLMs to simulate survey respondents

MARCH 22, 2024 | 5 MIN READ

Can AI Replace Human Research Participants? These Scientists See Risks

Several recent proposals for using AI to generate research data could save time and effort but at a cost

BY CHRIS STOKEL-WALKER

Synthetic respondents are the homoeopathy of market research

Published on
11 March 2024

Share

Nik Samoylov
Director

AI polling company defends wrong predictions on the US election

Diego Mendoza

Nov 6, 2024, 9:26pm GMT+1 TECH POLITICS NORTH AMERICA

Bad Idea!? | Use LLMs to simulate survey respondents

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

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

Whose Opinions Do Language Models Reflect?

Shibani Santurkar [1] Esin Durmus [1] Faisal Ladhak [2] Cinoo Lee [1] Percy Liang [1] Tatsunori Hashimoto [1]

Questioning the Survey Responses of Large Language Models

Ricardo Dominguez-Olmedo Moritz Hardt Celestine Mendler-Dünner

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Problem | Generalizability?

- **Biased LLM output:** stereotypes, political attitudes, WEIRD* perspectives
- One potential reason: **unrepresentative training data**
  - prevalence of native-**language** training data
  - **political and social** structure & public opinion dynamics
  - **digital divide:** target population ↔ **population reflected** in training data

→ challenges validity
→ risks reinforcing biases in research, politics, society
➔ Need to test LLM-synthetic samples in different contexts

*Western, Educated, Industrialized, Rich, Democratic

Research Gaps

- Comparative studies based on country-level prompting vs. individual-level prompting only single-country studies
- Biases related to prompt language or content?
- “Predicting the past” vs. future outcomes

➔ Test LLMs’ predictive performance …

→ across *national and linguistic* contexts based on *individual-level* prompts

→ with *limited individual-level information* (feasibility of repurposing survey data)

→ for *future* outcomes

Research Questions

→ Can LLMs predict the aggregate results of *future* elections?

→ How does LLMs' predictive performance differ across *countries* and *languages*?

→ How does LLMs' predictive performance differ depending on the *information provided* in the prompt?

→ Are there differences in performance between LLMs?

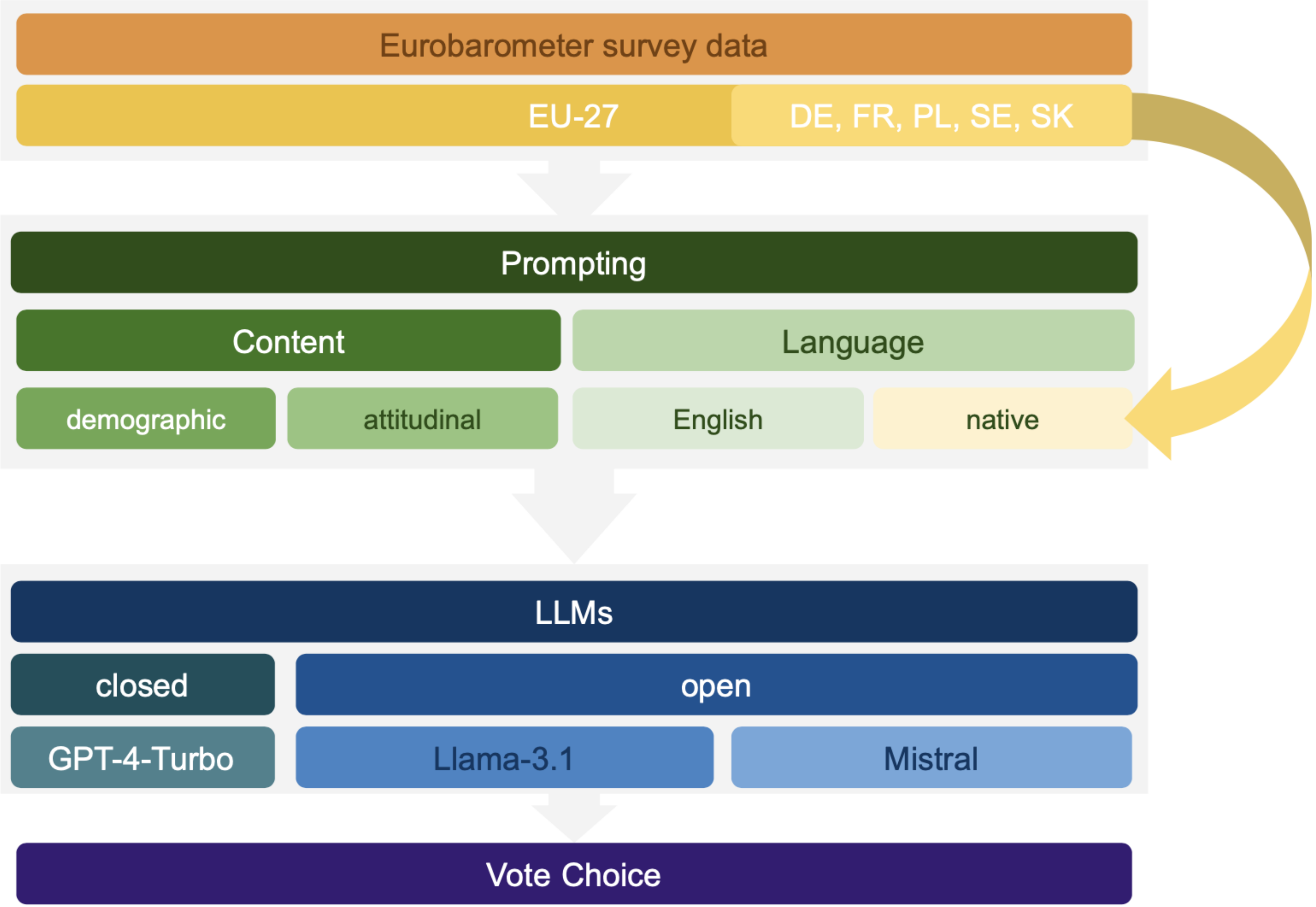
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Test Case | Vote choice in the 2024 EU elections

- **Vote choice** – popular item in public opinion research:
    - real-world relevance
    - challenging to predict → with vs. without LLMs?
    - much-discussed in (online) research & society → covered by training data?
    - correlates with factors potentially limiting generalizability of U.S.-based findings
- **EU elections**
    - covering several different populations, party systems, languages, …
    - future outcome at time of data collection

Research Design | Experimental dimensions

1. Country
2. Prompt Content
3. Prompt Language
4. LLM



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Research Design | Data

Create personas based on survey data

Prompt LLMs with personas

Compare predictions to election results

Countries	EU-27; DE, FR, IE, PL, SE, SK
Prompt Languages	English; German, French, Polish, Swedish, Slovak
Dataset	Eurobarometer 99.4 (May/June 2023)
Sample	Voting-eligible participants living in EU member states (n=26,047)
Prompt Variables	**Demographics:** age, gender, education, occupation, class, urbanity **Attitudes:** political interest, ideological left-right self-placement, attitude towards EU integration, trust in EU

Research Design | Prompt design

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Create personas based on survey data

Prompt LLMs with personas

Compare predictions to election results

	country <chr>	age <dbl>	gender <fctr>	education <fctr>	class <fctr>	occupation <fctr>	urbanity <fctr>	ideology <fctr>	interest <fctr>	trust_eu <fctr>
1	Austria	48	female	secondary	lower middle	working	large city	center-left	not at all	not to trust the European Union
2	Austria	41	female	university	middle	working	rural area	center-left	moderately	not to trust the European Union
3	Austria	57	male	vocational	working	working	large city	in the middle	very	to trust the European Union
4	Austria	81	female	vocational	middle	retired	large city	strongly left	very	to trust the European Union
5	Austria	43	female	university	middle	working	large city	strongly right	very	to trust the European Union
6	Austria	26	female	vocational	lower middle	working	large city	strongly right	moderately	to trust the European Union

```r
    prompt_en_full = paste0("The year is 2024. You are a voting-eligible citizen of an EU-member state
living in ", country,
                       ". The parties competing in the 2024 elections to the European Parliament are ",
parties_country_random,
                       ", and several smaller parties. You are ", age,
                       " years old. You are ", gender,
                       ". You have a ", education,
                       " education. Socio-economically, you are ", class,
                       " class. You are ", occupation,
                       ". You live in a ", urbanity,
                       ". You are ", interest,
                       " interested in politics. Ideologically, you are ", ideology,
                       ". You ", integration_decisions,
                       ". You tend ", trust_eu,
                       ". Will you vote in the 2024 elections to the European Parliament, and if so, for
which party? Use the information above to make as good a prediction as possible, and keep your answer as
short as possible, if possible only 'No' or the name of the party you would predict."),
```

prompt_en_full <chr>
The year is 2024. You are a voting-eligible citizen of an EU-member state living in Austria. The parties competing in the 2024 elections to the Euro...
The year is 2024. You are a voting-eligible citizen of an EU-member state living in Austria. The parties competing in the 2024 elections to the Euro...
The year is 2024. You are a voting-eligible citizen of an EU-member state living in Austria. The parties competing in the 2024 elections to the Euro...
The year is 2024. You are a voting-eligible citizen of an EU-member state living in Austria. The parties competing in the 2024 elections to the Euro...
The year is 2024. You are a voting-eligible citizen of an EU-member state living in Austria. The parties competing in the 2024 elections to the Euro...
The year is 2024. You are a voting-eligible citizen of an EU-member state living in Austria. The parties competing in the 2024 elections to the Euro...

Research Design | Prompt design

Create personas based on survey data

Prompt LLMs with personas

Compare predictions to election results

The year is 2024. You are a voting-eligible citizen of an EU member state living in **Germany**. The parties competing in the 2024 elections to the European Parliament are **CDU/CSU, SPD, Grüne, FDP, Linke, AfD, Freie Wähler, BSW, Volt,** and several smaller parties. You are **28** years old. You are **female**. You have a **university** education. Economically, you are **upper-middle** class. You are **working**. You live in a **big city**. <u>You are **very** interested in politics. Ideologically, you are **center-left**. You **think that** more decisions should be taken at the EU-level. You tend **to trust** the European Union.</u> Will you vote in the 2024 elections to the European parliament, and if so, for which party? Use the information above to make as good a prediction as possible, and keep your answer as short as possible, if possible only "No" or the name of the party you would predict.

*Example prompt. Variables **bold**. Attitudinal information <u>underlined</u>.*

Research Design | LLM configuration

Create personas based on survey data

Prompt LLMs with personas

Compare predictions to election results

```
output_en_at_full <- rgpt( # rename for distinguishing datasets later on
  prompt_role_var = EB994_EN_AT$role, # adjust df
  prompt_content_var = EB994_EN_AT$prompt_en_full, # adjust df and column
  param_seed = 240524,
  id_var = EB994_EN_AT$uniqid, # adjust df
  param_output_type = "complete",
  param_model = "gpt-4-turbo",
  param_max_tokens = 40,
  param_temperature = 0.9,
  #defaults / not using:
  param_top_p = 1,
  param_n = 1,
  param_stop = NULL,
  param_presence_penalty = 0,
  param_frequency_penalty = 0
)

completions_en_at_full <- output_en_at_full[[1]] # extract completions

metadata_en_at_full<- output_en_at_full[[2]] # extract metadata
```

Kleinberg, B. (2024). *rgpt3: Making requests from R to the GPT API* (Version 1.0) [Computer software]. https://doi.org/10.5281/zenodo.7327667

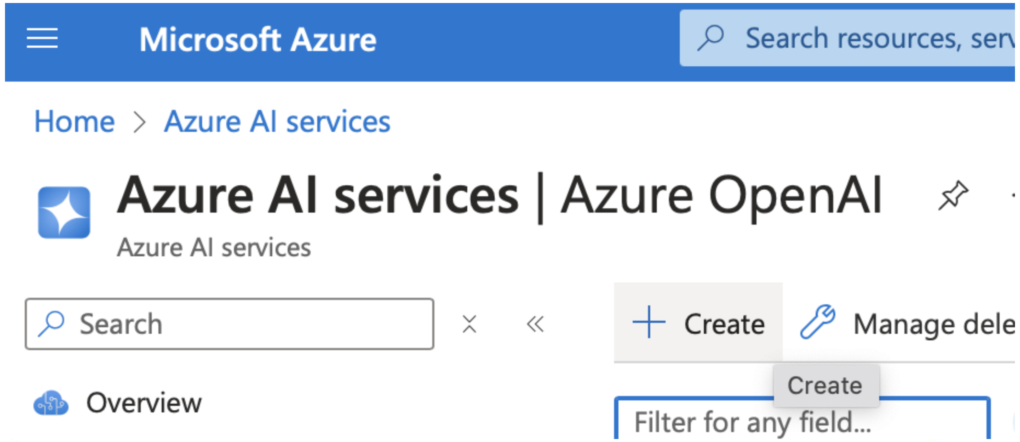
Data collection: June 2024

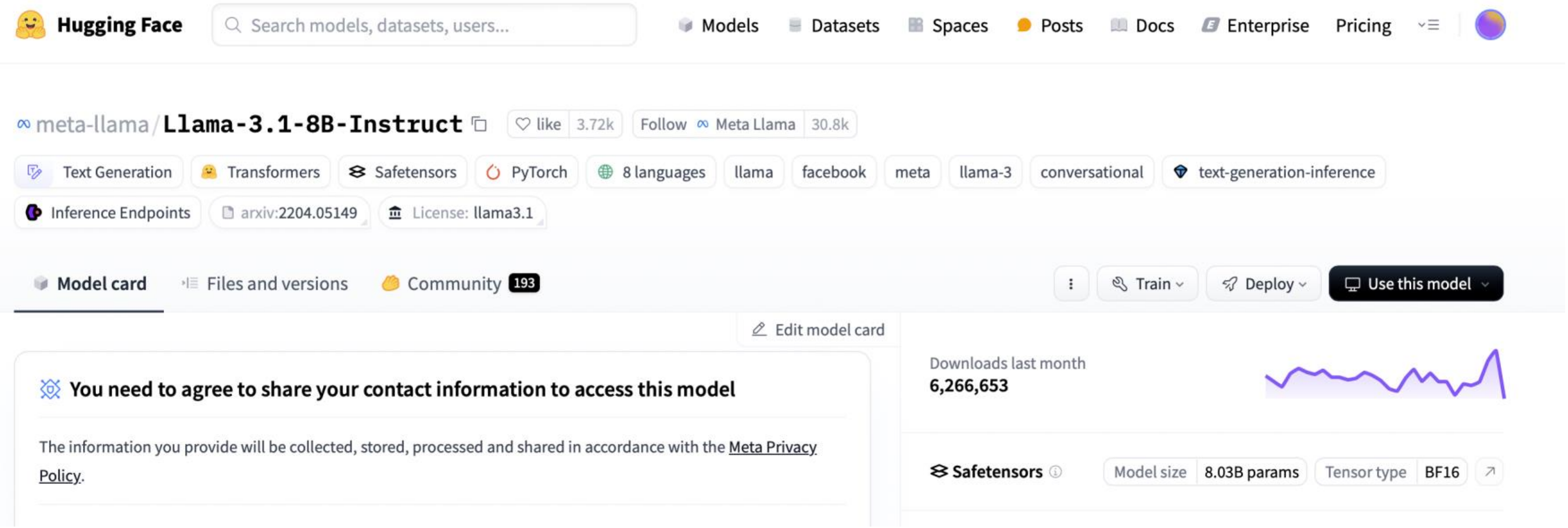
Research Design | LLM configuration

Create personas based on survey data

Prompt LLMs with personas

Compare predictions to election results





Research Design | Analysis

Create personas based on survey data

Prompt LLMs with personas

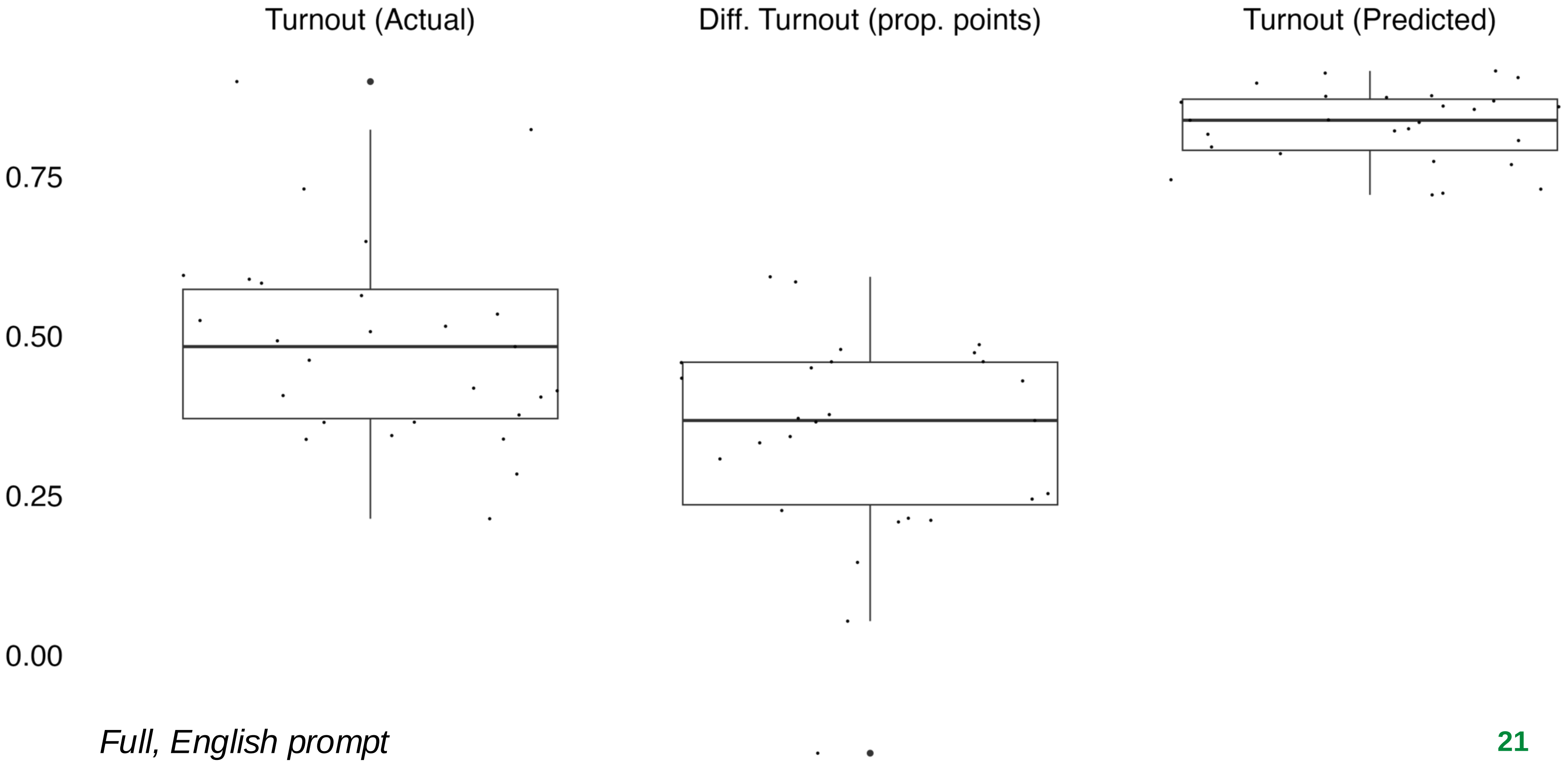
Compare predictions to election results

- Weight output with survey weights
- Aggregate per-country analysis:
  difference between prediction and election results
- Distinguish turnout vs. party vote shares
- Dimensions of comparison:
  - **Societal coverage → countries:** region (social & political contexts, digital divide), language family
  - **Linguistic coverage → prompt language:** English vs. native language
  - **Attitudinal coverage → prompt content:** Demographic information only vs. added attitudinal information

Results | Can LLMs predict the aggregate results of **future** elections?

Turnout

- predicted (avg.): 83%
- actual (avg.): 49%; higher variation

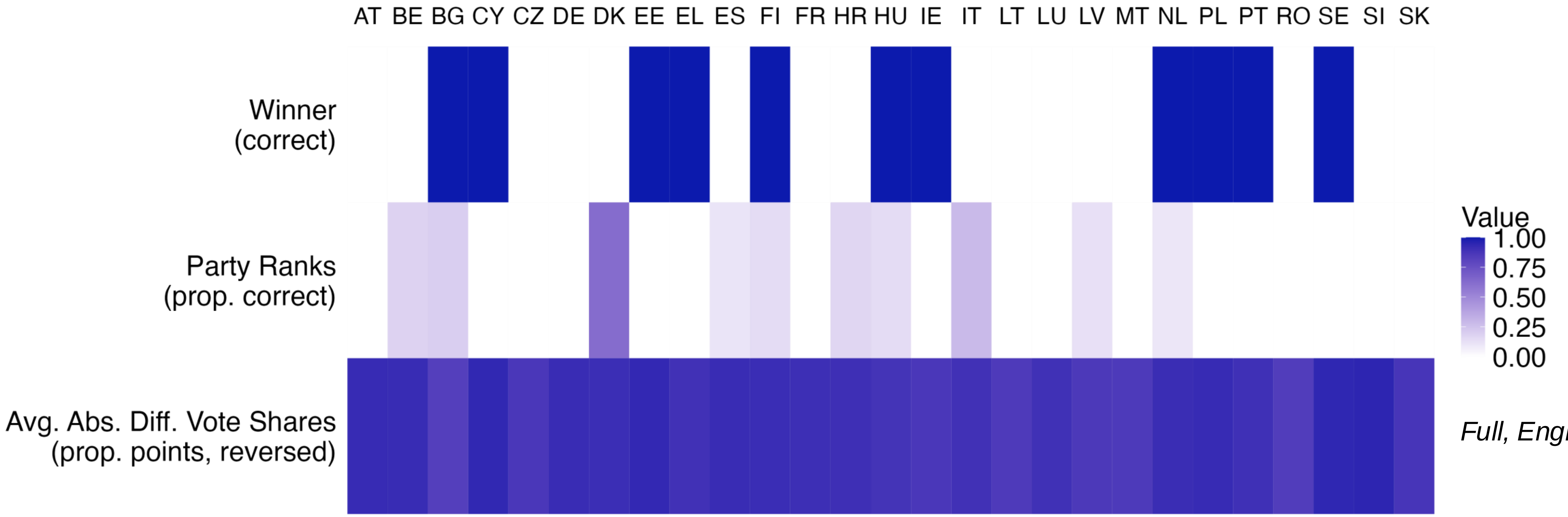


Full, English prompt

Results | Can LLMs predict the aggregate results of **future** elections?

Party vote shares

- 11/27 winners correct
- avg. ranks correct: 8% (median: 0)
- avg. differences: 7-15 percentage points



Full, English prompt

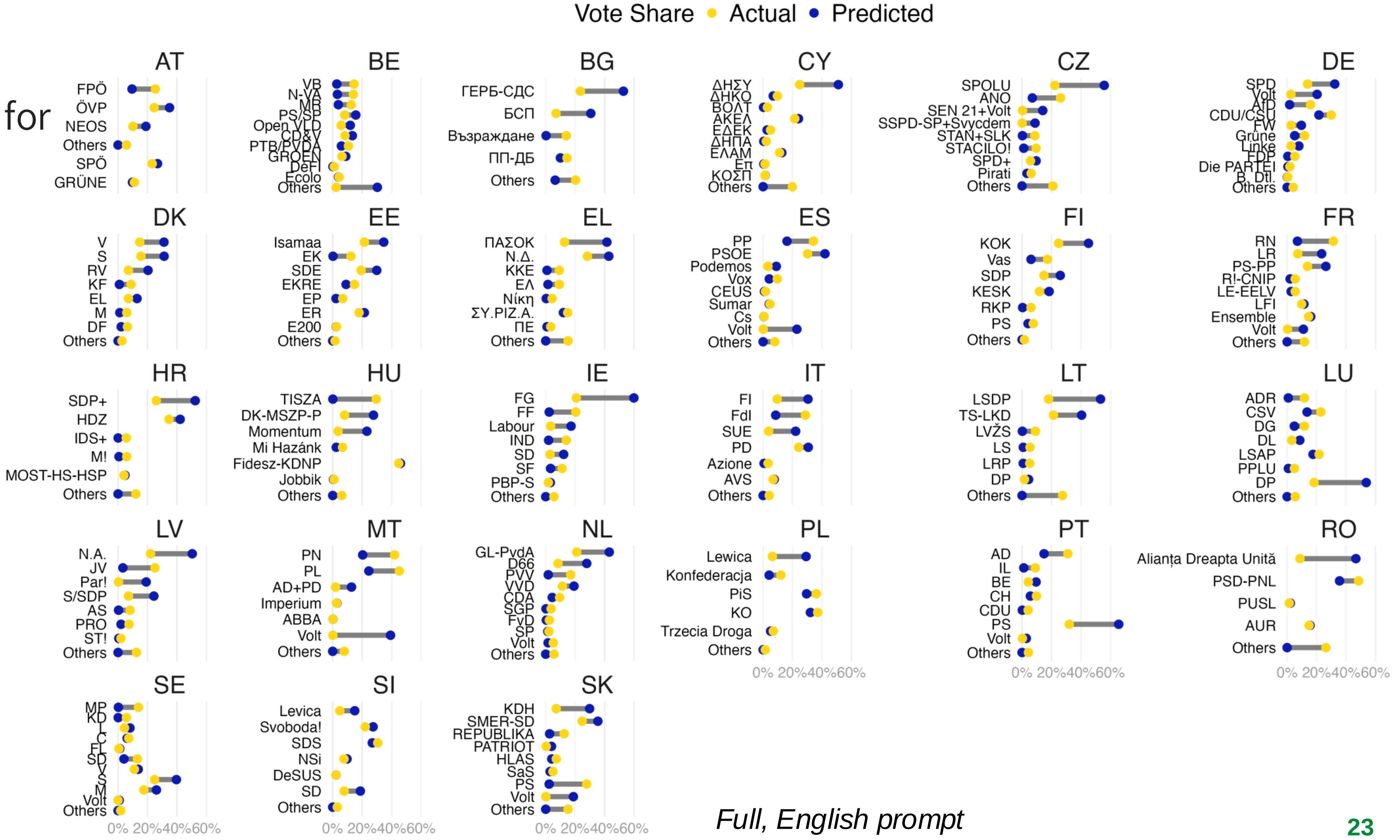
Note: Average absolute differences in vote shares: higher values correspond to better predictive performance.
Example: an average absolute difference of 5 percentage points (0.05) would be displayed as 0.95.

Results | Can LLMs predict the aggregate results of **future** elections?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Party vote shares

- larger differences for non-green or -left parties

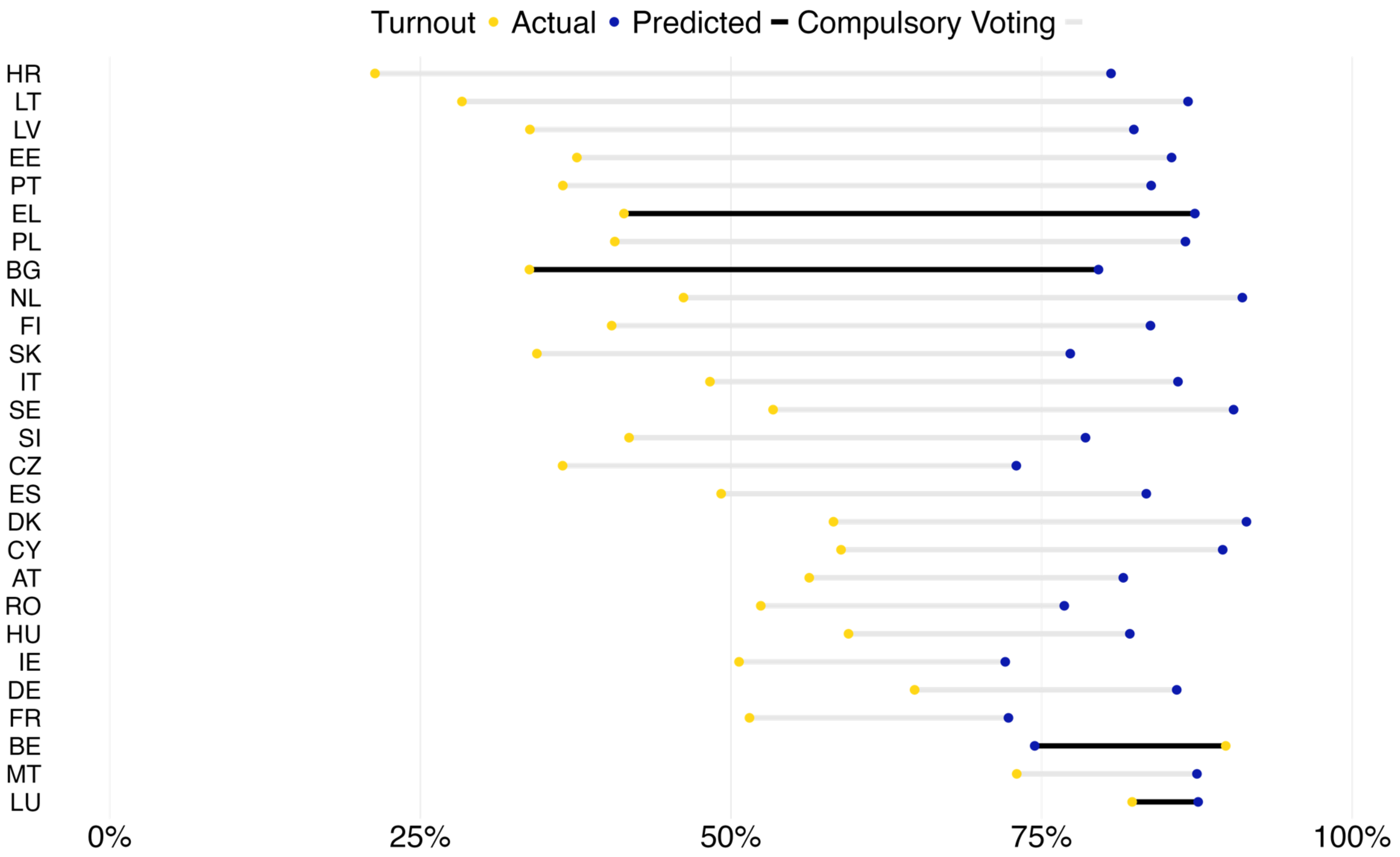


Full, English prompt

Results | How does predictive performance differ across **countries**?

Turnout

- better for countries with high actual turnout
- compulsory voting not relevant for predictions



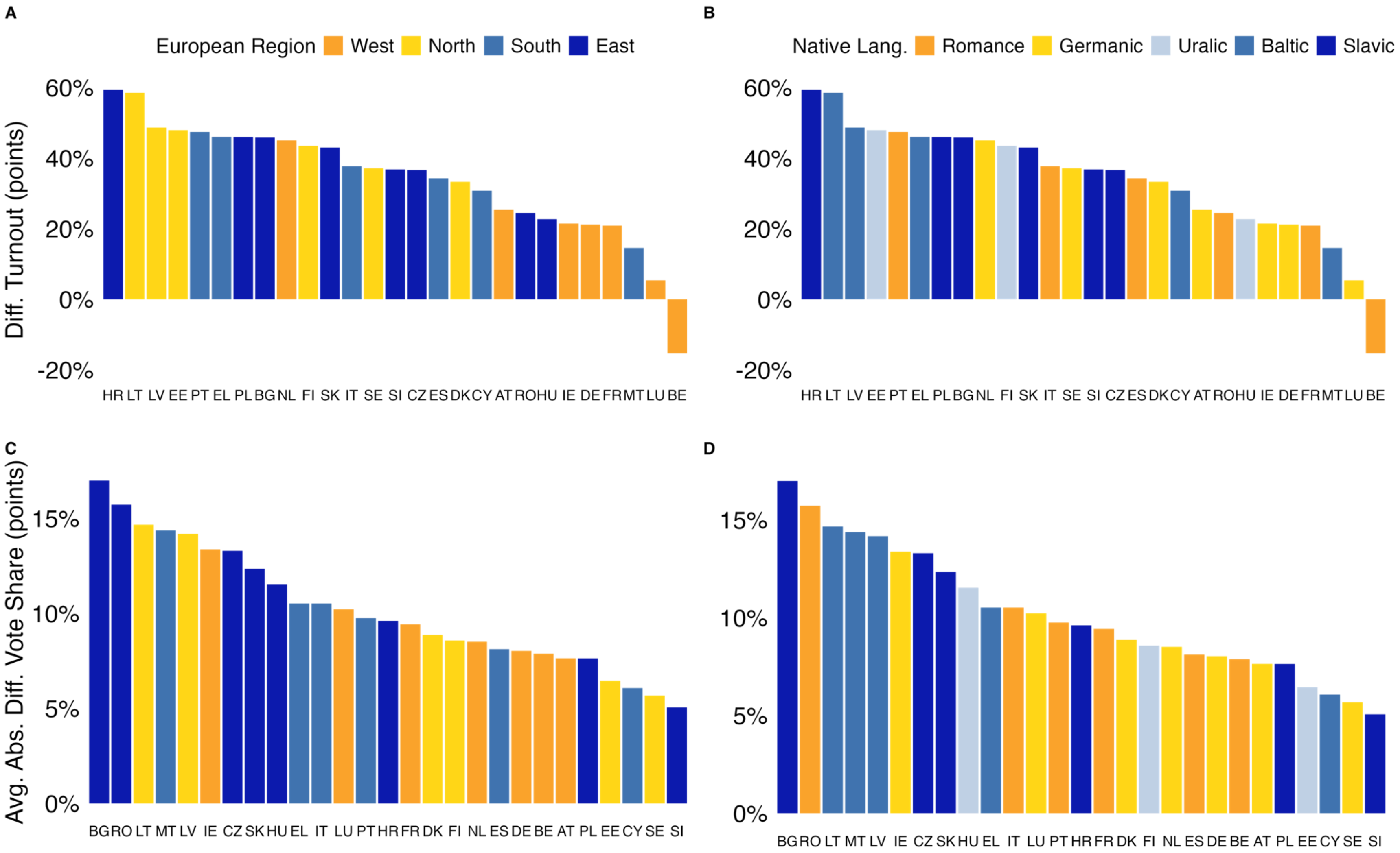
Full, English prompt

Results | How does predictive performance differ across **countries**?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Turnout & party vote shares

- better for Western countries with more dominant languages
- worse for Eastern European countries with Slavic languages



Full, English prompt

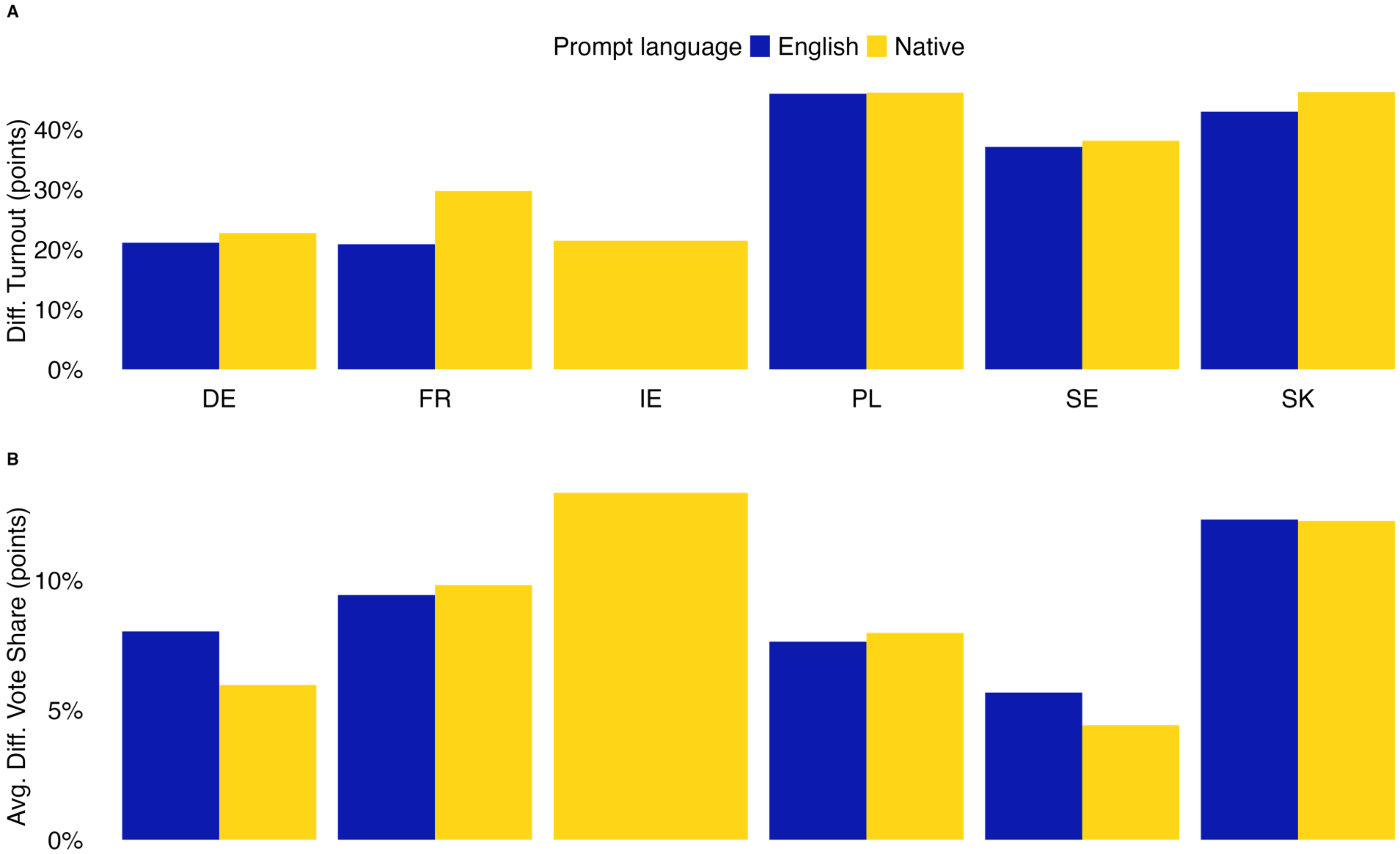
Results | How does predictive performance differ across **languages**?

Turnout

- worse when prompted in native language
- no difference (already bad) in PL

Party vote shares

- better when prompted in English (DE, SE)
- slightly worse for FR, PL

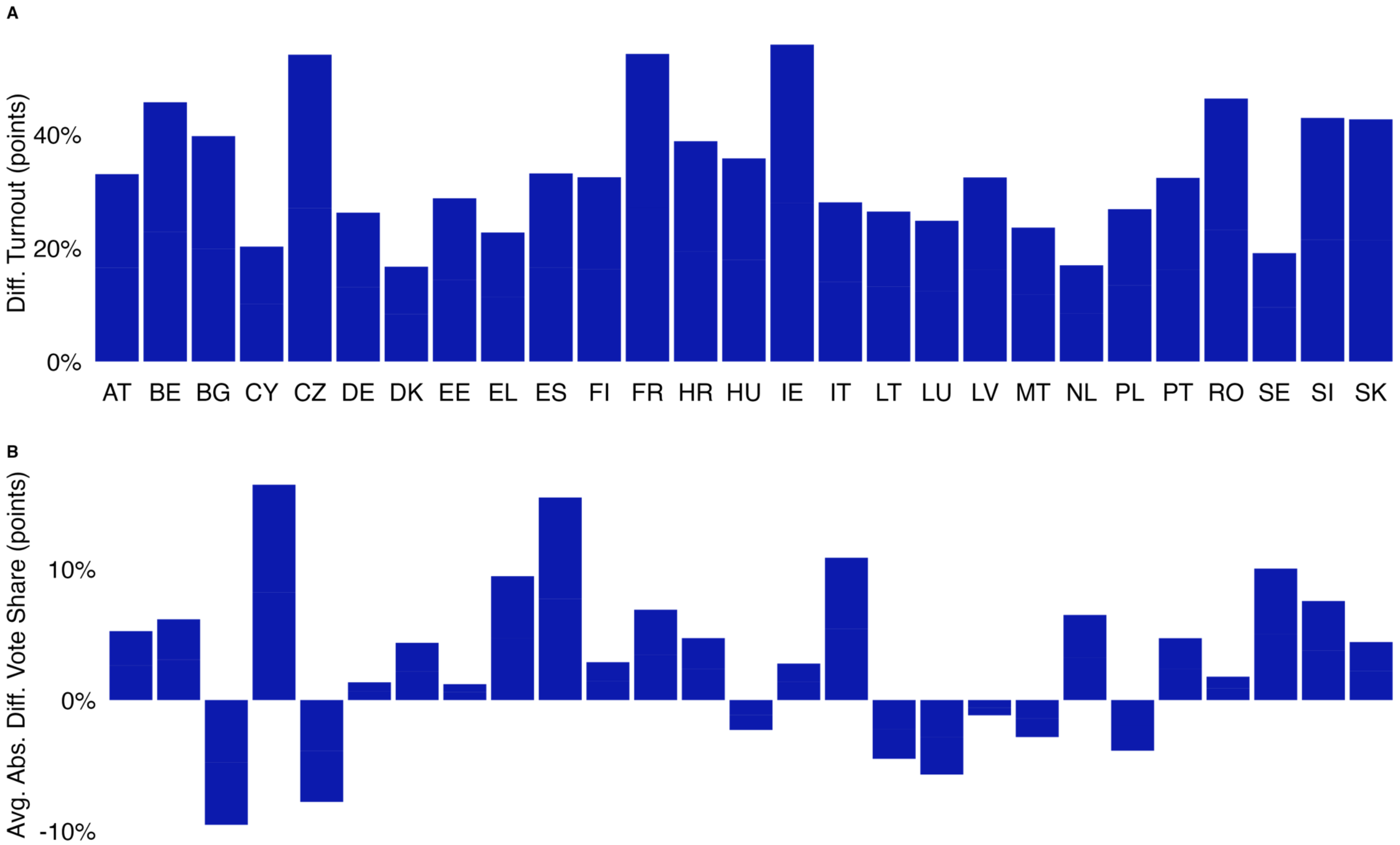


Full prompt

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Results | Does predictive performance depend on **information** in the prompt?

Turnout & party vote shares
- even worse with only demographic information
  - regardless of prompt language
- lower variance in vote share differences
  → systematically off?



Difference demographics only vs. full English prompt

Results | Are these problems LLM-specific?

LLaMa 3.1: similar patterns as GPT-4-Turbo

- **Overall/Country:** Even higher overestimations and bigger biases (again Eastern European / Slavic countries) for turnout, smaller for vote shares → bias generalizable
- **Prompt language:** Even poorer predictive performance with native language prompt → limited multilingual capacities
- **Prompt content:** Even worse predictions with demographic-only prompt
- Higher shares of missing predictions

Mistral 7B: unable to complete task

- “Difficult to say with certainty”
- Not following instruction to keep answer concise → responses cut off
- More missing predictions with demographic-only prompt

Summary | Just because you can, doesn't mean you should

… but can you even?

LLM-based predictions of aggregate results of the 2024 European elections **fail**:

- overestimate turnout
- unable to accurately predict the winner, rank ordering, or individual party vote shares
- especially off for **Eastern European** countries and countries with native **Slavic** languages
- especially off given only socio-demographic **information** about individual voters

Summary | Just because you can, doesn't mean you should

… but can you even? → Possible improvements:

- considering country-specific factors in prompting: prompt variables associated with vote choice (if available in survey data)
- building more sophisticated forecasting models (likely voters ?)
- using pre- & post-election panel as baseline

→ **secondary** data not available **pre-election**!

- considering country-specific factors in forecasting:
  - electoral systems & thresholds
  - party system fragmentation
  - electoral volatility
  - strategic voting

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Implications | Survey’s ain’t dead yet

- (*General*-purpose / off-the-shelf) **LLMs were not made** for predicting *specific* public opinion!
- Performance of LLMs is dependent on **training data and prompt**

    → **Training data** temporality:

        → Volatility of population structure & attitudes

        → Tradeoff between recency and detail of human samples needed for personas

        → Training data cutoffs

    → **Prompt:** Need detailed attitudinal information to make somewhat more accurate predictions

→ Questionable feasibility of using LLM-based synthetic samples as a supplement or substitution of detailed survey data!

Conclusion | What's next?

Needs:

- **Bias identification & mitigation:** Transparency & diversity in model architectures & training data
- **Purpose optimization:** Customizing LLMs for
  - public opinion estimation
  - underrepresented contexts

HUMAN PREFERENCES IN LARGE LANGUAGE MODEL LATENT SPACE: A TECHNICAL ANALYSIS ON THE RELIABILITY OF SYNTHETIC DATA IN VOTING OUTCOME PREDICTION

Sarah Ball*[1,5], Simeon Allmendinger*[2,4], Frauke Kreuter[1,3,5], and Niklas Kühl[2,4]

Fine-Tuning Large Language Models to Simulate German Voting Behaviour (Working Paper)

Tobias Holtdirk[1], Dennis Assenmacher[1], Arnim Bleier[1], Claudia Wagner[1,2]
[1]GESIS - Leibniz Institute for the Social Sciences
[2]RWTH Aachen
{firstname.lastname}@gesis.org

Article | Open access | Published: 05 June 2024

Scaling neural machine translation to 200 languages

NLLB Team

Nature **630**, 841–846 (2024) | Cite this article

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

Key Takeaways

As of now …

- ➢ LLMs **cannot replace survey data** (at most augment it)

- ➢ Applicability of LLM-generated survey data is **context-dependent**
  → output is biased towards certain (sub-)populations

- ➢ Performance likely improves with **fine-tuning**

- ➢ More research needed for **identifying & mitigating LLM biases**

Questions? Collaborations?
Let’s connect!

Leah von der Heyde
L.Heyde@lmu.de
linkedin.com/in/leahvonderheyde

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Selected references

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). *The illusion of artificial inclusion*. https://doi.org/10.1145/3613904.3642703

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*, 1–15. https://doi.org/10.1017/pan.2023.2

Ball, S., Allmendinger, S., Kreuter, F., & Kühl, N. (2025). *Human Preferences in Large Language Model Latent Space: A Technical Analysis on the Reliability of Synthetic Data in Voting Outcome Prediction* (No. arXiv:2502.16280). arXiv. https://doi.org/10.48550/arXiv.2502.16280

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 1–16. https://doi.org/10.1017/pan.2024.5

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2023). *Questioning the Survey Responses of Large Language Models* (No. arXiv:2306.07951). arXiv. http://arxiv.org/abs/2306.07951

Kim, J., & Lee, B. (2023). *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction* (No. arXiv:2305.09620). arXiv. http://arxiv.org/abs/2305.09620

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve* (No. arXiv:2309.13638). arXiv. http://arxiv.org/abs/2309.13638

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? *Proceedings of the 40th International Conference on Machine Learning*, 29971–30004. https://proceedings.mlr.press/v202/santurkar23a.html

Selected references & further resources

Mendoza, D. (2024, November 6). AI polling company defends wrong predictions on the US election. *Semafor*. https://www.semafor.com/article/11/06/2024/ai-startup-aaru-defends-using-artificial-intelligence-for-polling

NLLB Team, Costa-jussà, M. R., Cross, J., Celebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., … Wang, J. (2024). Scaling neural machine translation to 200 languages. *Nature*, *630*(8018), 841–846. https://doi.org/10.1038/s41586-024-07335-x

TrustLLM. (n.d.). *TrustLLM: Democratizing Trustworthy and Factual Large Language Model Technology for Europe*. TrustLLM. Retrieved August 21, 2024, from https://trustllm.eu/

- Preprint with partial results from this study: https://doi.org/10.48550/arXiv.2409.09045
- Previous work with subgroup-level analysis and comparison to survey-reported vote choice: https://doi.org/10.48550/arXiv.2407.08563
- Related literature (non-comprehensive/systematic)
  - https://github.com/Value4AI/Awesome-LLM-in-Social-Science
  - https://github.com/penguinnnnn/awesome-llm-and-society