1

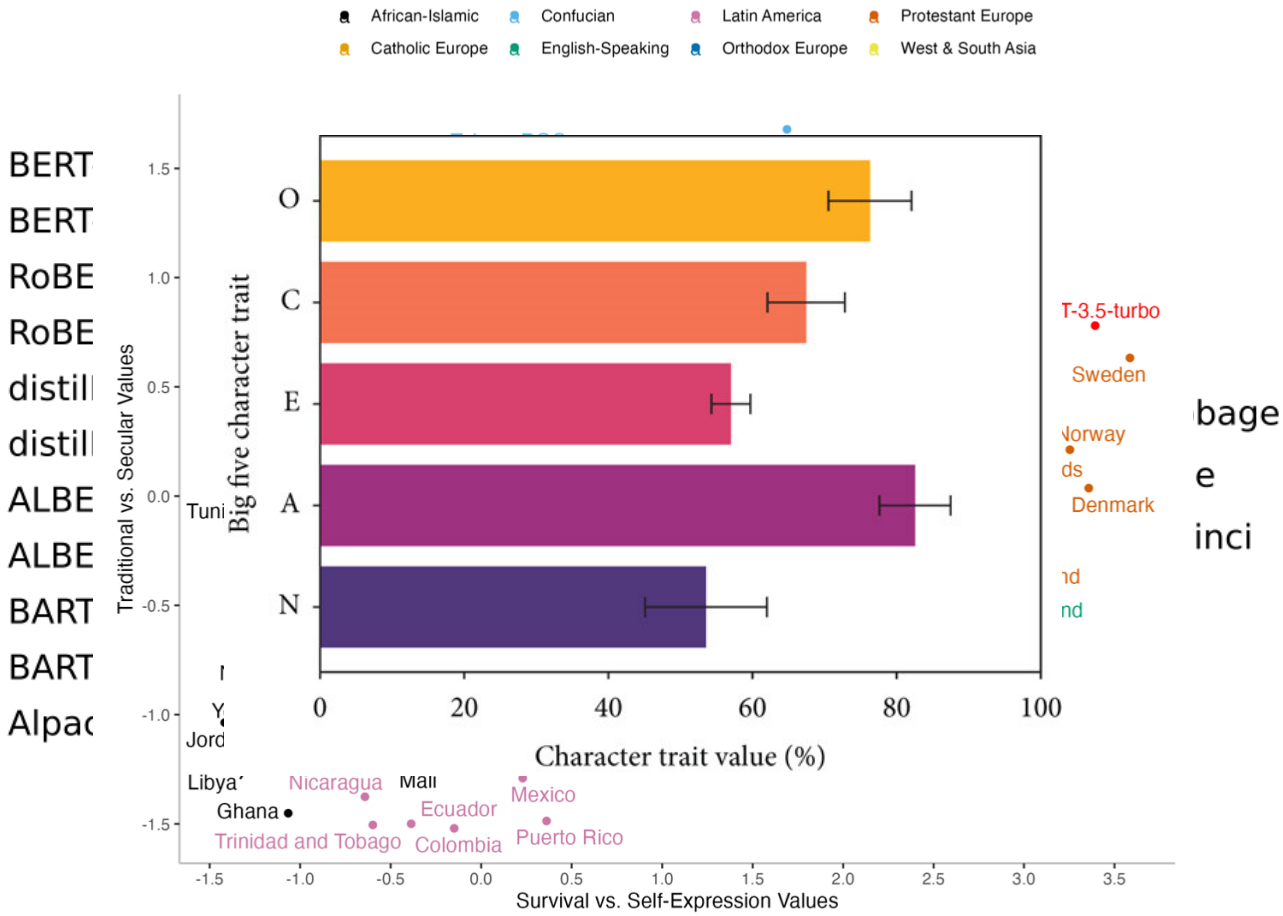
Exploring Large Language Models Political Biases

Noé Durandard
ENS-PSL, Lattice

November 25, 2024

AI for science, science for AI

3



Political Compass Test's results of various LMs [1], The Inglehart-Welzel World Cultural Map with GPT [2], ChatGPT's Big-5 test results [3].

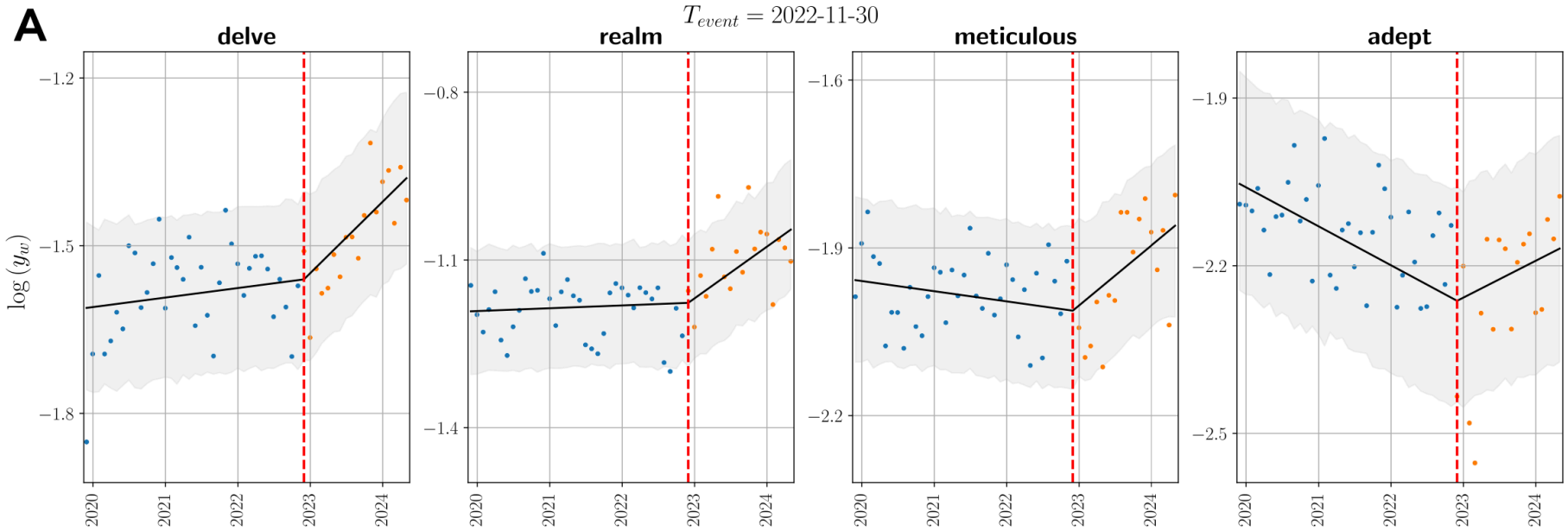
5

1 Real-world implications

6

1.1 Evidences of ChatGPT's impact

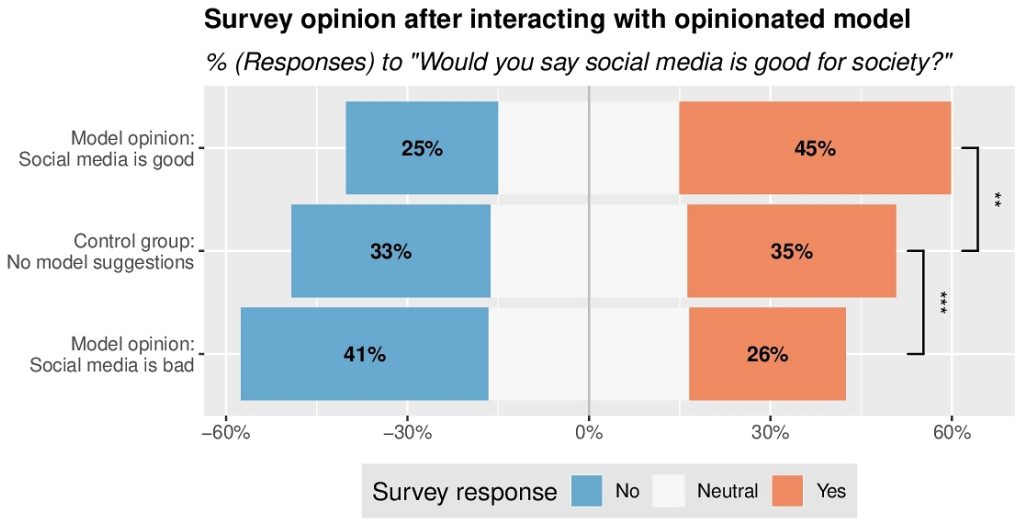
LLMs influence the way we communicate [4–6].



Prevalence of words associated with ChatGPT. Extracted from [4].

7

1.2 Influence besides vocabulary



Participants interacting with a model supportive of social media were more likely to say that social media is good for society in a later survey (and vice versa) — from [7].

LLMs influence users and participate in shaping opinion

- collective opinion [8]
- individual opinion [7, 9]

8

1.3 Propagation of models' biases

Language is loaded with sociocultural characteristics [10].

LLMs biases can have harmful consequences, and spread in downstream applications:

- lower quality service to minorities and vulnerablue users [11, 12]
- demographic-based recommandations [13]
- mental health issues and cultural values [14]
- politically biased text summarization [15]
- …

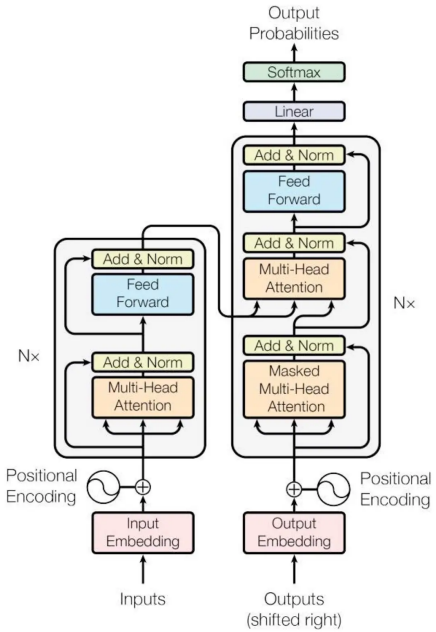
10

2 LLMs Crash course

11

2.1 Overview

- Trasnformers-based architecture [16]
- Autoregressive models
- Trained over **large** corpora (typically 10^{11} - 10^{12} tokens) [e.g. 17]
- Further *aligned* with desired behaviors



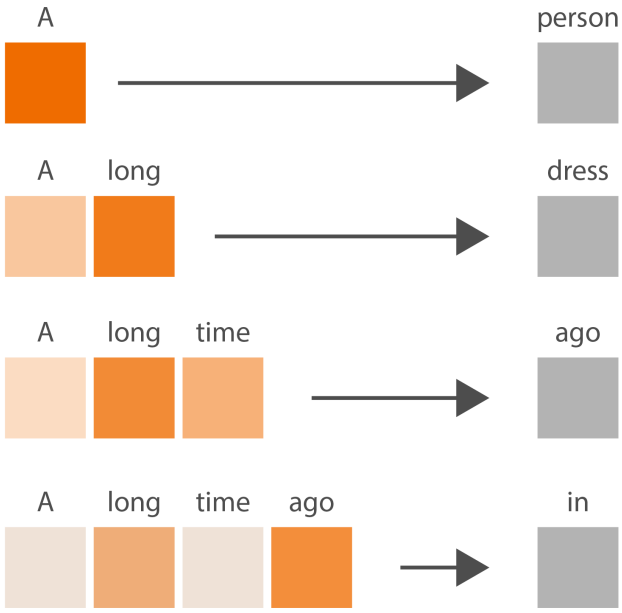
Transformers architecture from [16].

12

2.2 Next-token prediction

$$\mathcal{L}[\phi] = -\sum_{i=1}^{I}\sum_{t=1}^{T}\log\left[Pr(x_{i,t+1}|x_{i,1},\cdots,x_{i,t},\phi)\right]$$

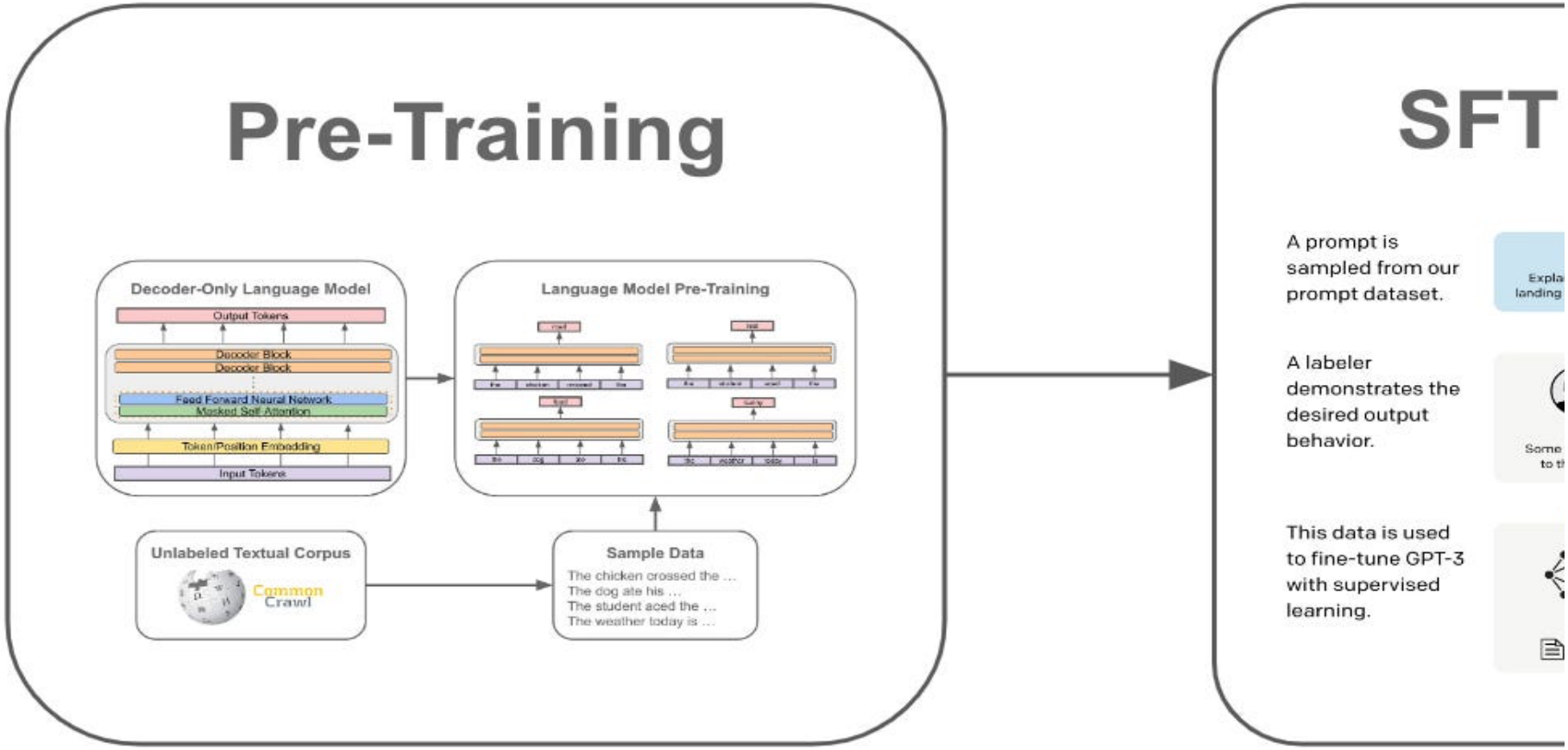
- language modeling task
- self-supervised training
- mainly webpages, code, books, encyclopedia
- equips models with other capabilities (world knowledge, comprehension, translation, …)



Next token prediction task. Extracted from [18]

13

2.3 LLMs training workflow



14

2.4 LLMs Evaluation

- Automated Evaluation
  - metric-based
  - objective assessment
- Human Evaluation
  - comaprative assessment
  - qualitative insights
  - vibe check
- Benchamrks datasets
  - diverse tasks and metrics
  - may rely on MCQA
- Adversarial Evaluation
  - robustness against attacks
  - identify and mitigate risks

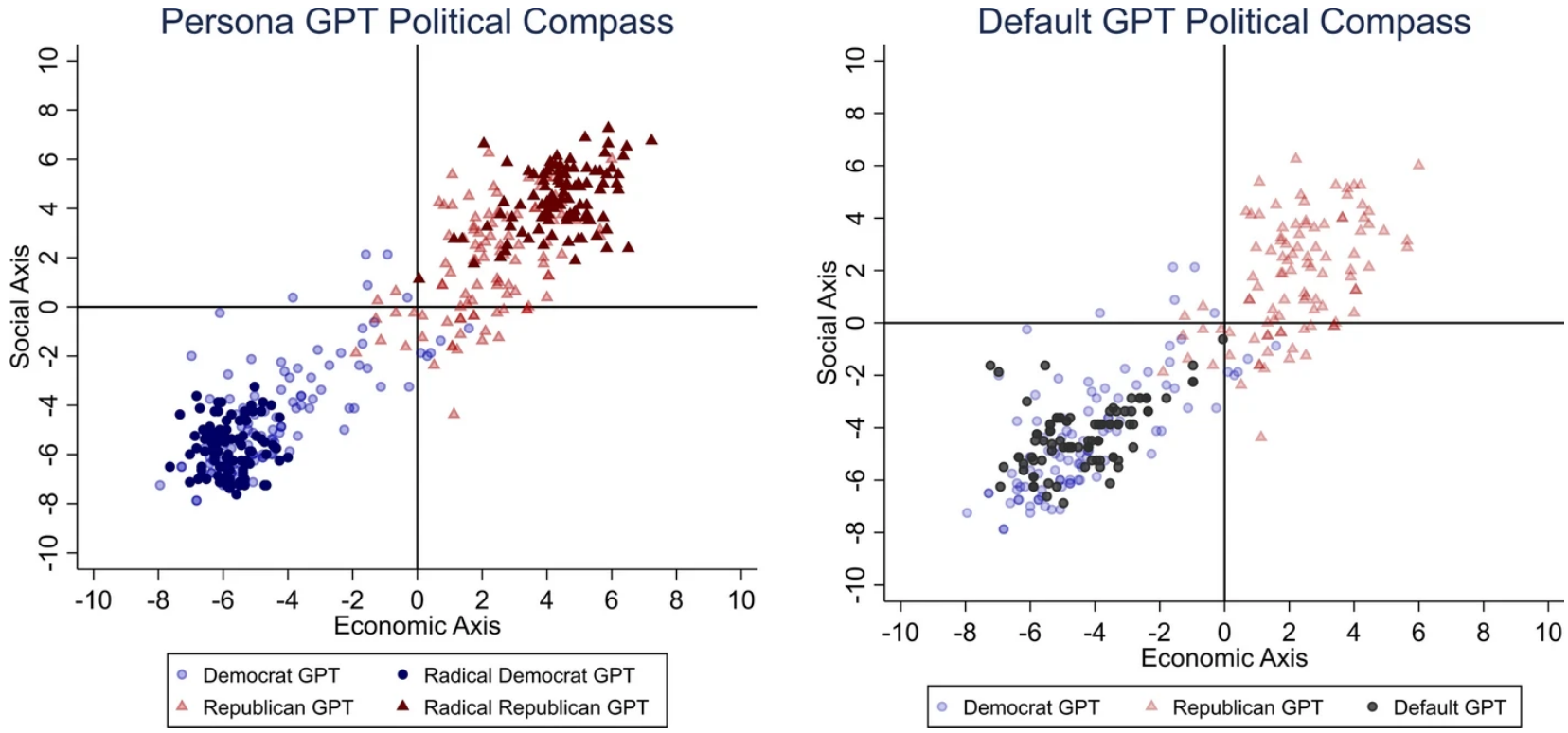
16

3 Political Biases in LLMs

17

3.1 The Political Compass

Many studies make use of the PCT [3, 19–21], highlighting biases towards progressive, left-leaning views.



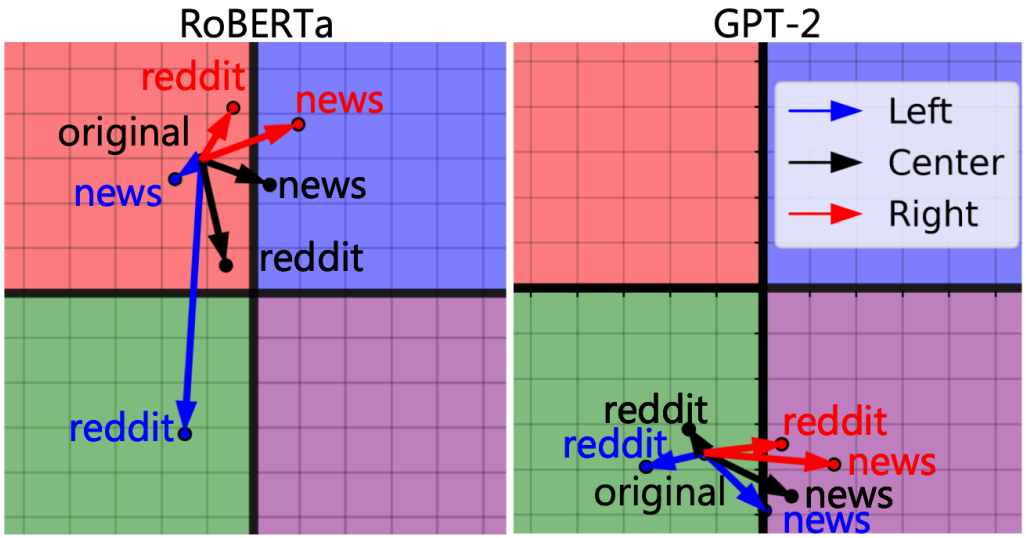
Extracted from [21].

18

3.2 Influencing behavior

Common mechanisms to modify LLMs behaviors:

- prompt engineering [20, 21]
- fine-tuning [1, 22]



PCT position after training on partisan corpora. From [1]

19

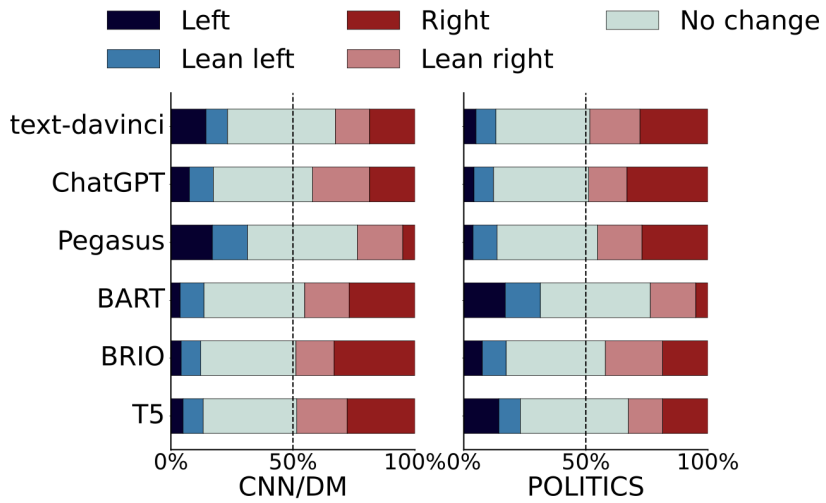
3.3 Methodological Issues

- Findings inconsistencies (large variance, prompt sensitivity [19], orthogonal context-dependent [23])
- Technical concerns
  - MCQA concerns [24, 25]
    - Probing methods discreapancies [26, 27]
    - Contamination or Clever Hans [28, 29]
    - Reflective Self-Assessment [30]
- **Not aligned with real-world use-cases** [19, 31]

20

3.4 Other frameworks?

- Political Stance: *what* is said and *how* it is said [32]
  - topical assessmment
- Summarization Task [15]
  - another task, another view?



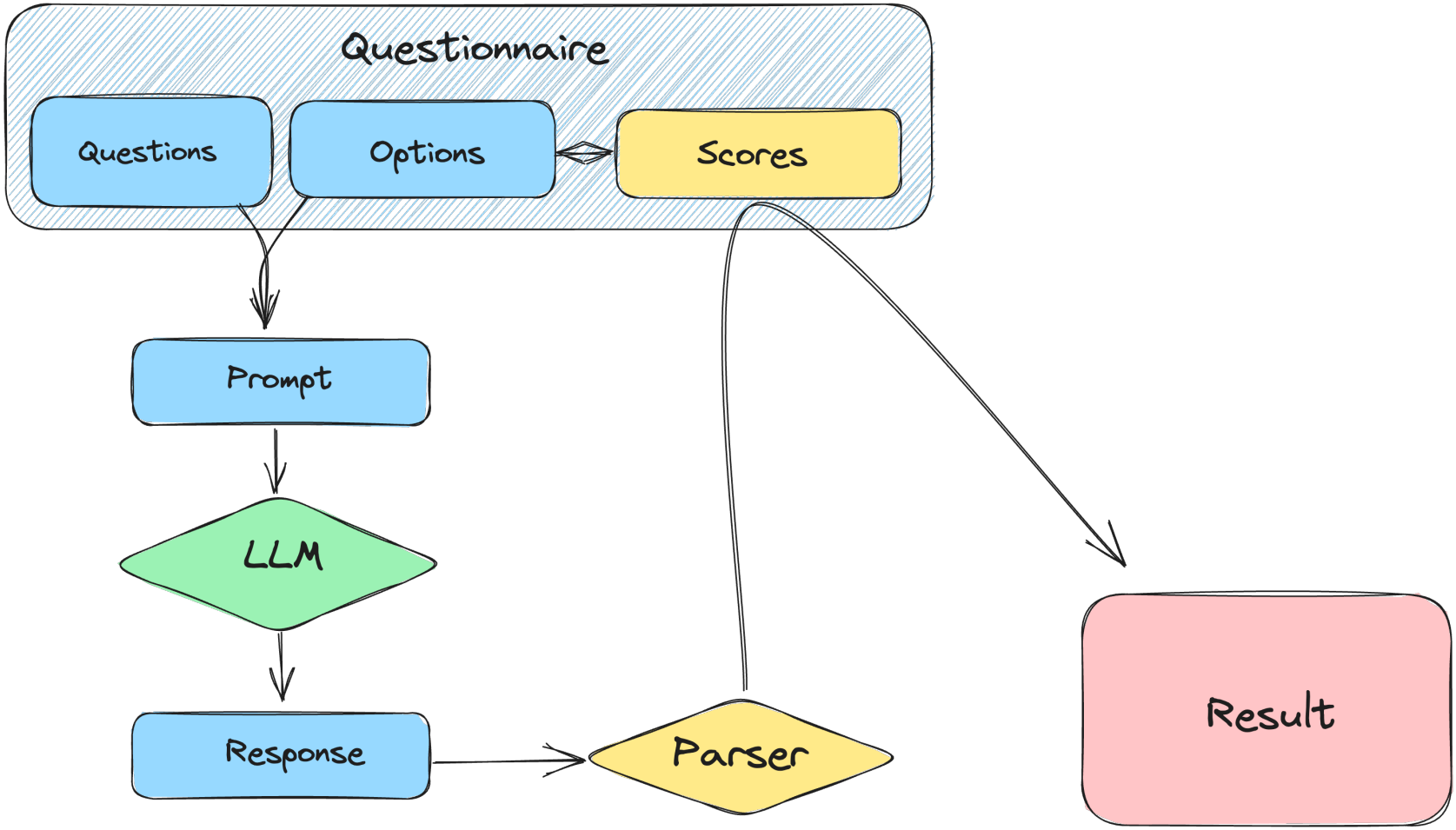
Changes in political stances between the summary and the article. From [15]

22

4 Project

23

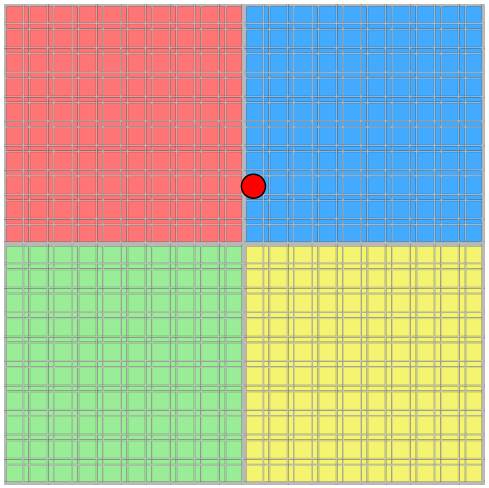
4.1 Base framework proposition



24

4.2 The Political Compass Test

Link to the original website



Economic 0.38 Social 2.41

Copy link to save or share answers

Reset Undo Redo Help

Purple libright Tiny indicator

Page 1

If economic globalisation is inevitable, it should primarily serve interests of trans-national corporations.

○ Strongly disagree ○ Disagree ◉ Agree ○ Strongly agr

I'd always support my country, whether it was ri

○ Strongly disagree ○ Disagree ◉ Agree ○ Strongly agr

No one chooses their country of birth, so it's foolish

○ Strongly disagree ○ Disagree ◉ Agree ○ Strongly agr

Our race has many superior qualities, compared w

○ Strongly disagree ○ Disagree ◉ Agree ○ Strongly agr

The enemy of my enemy is my frienc

○ Strongly disagree ○ Disagree ◉ Agree ○ Strongly agr

Military action that defies international law is some

○ Strongly disagree ○ Disagree ◉ Agree ○ Strongly agr

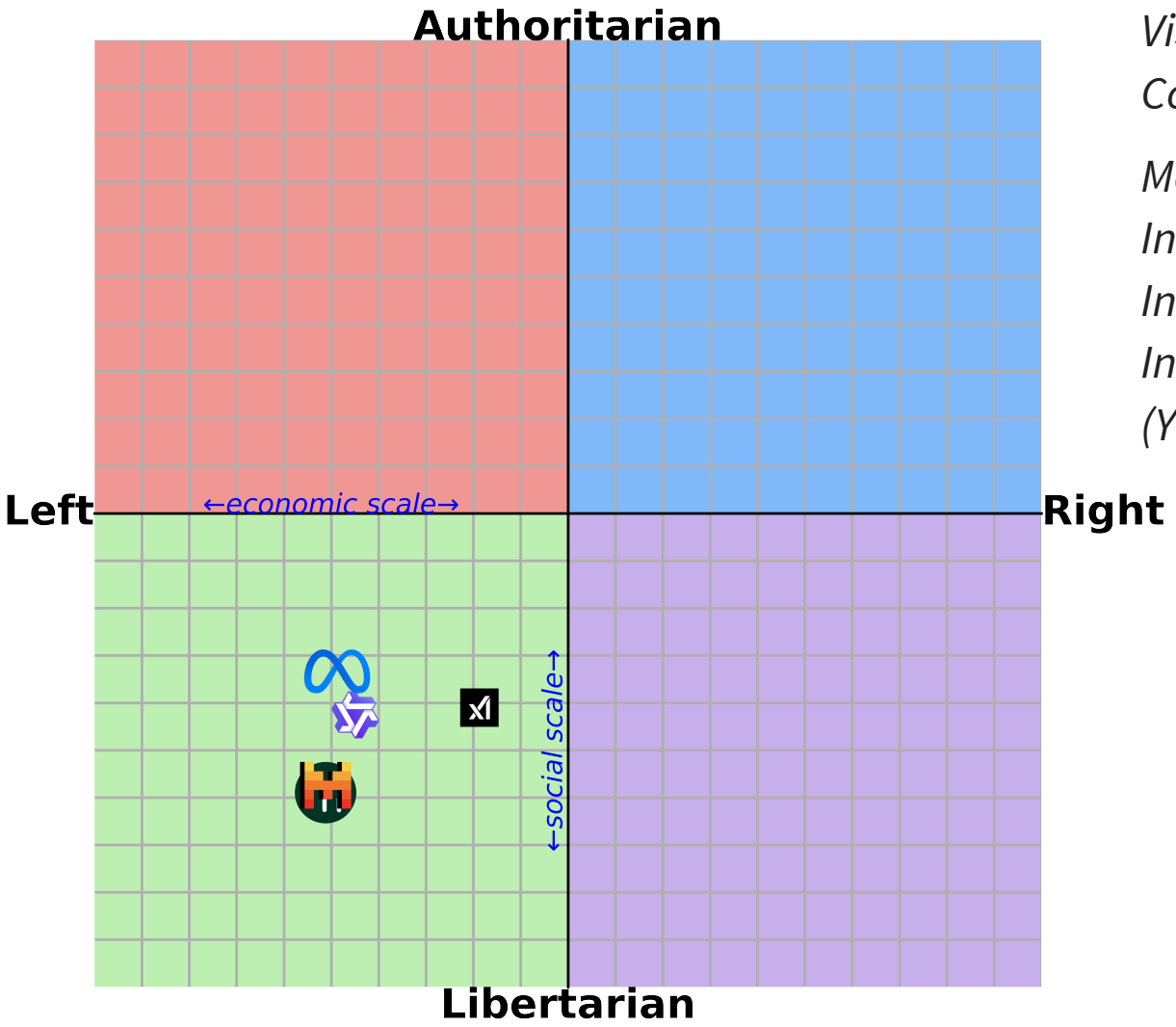
25

4.3 Code

GitHub link

26

4.4 Examples

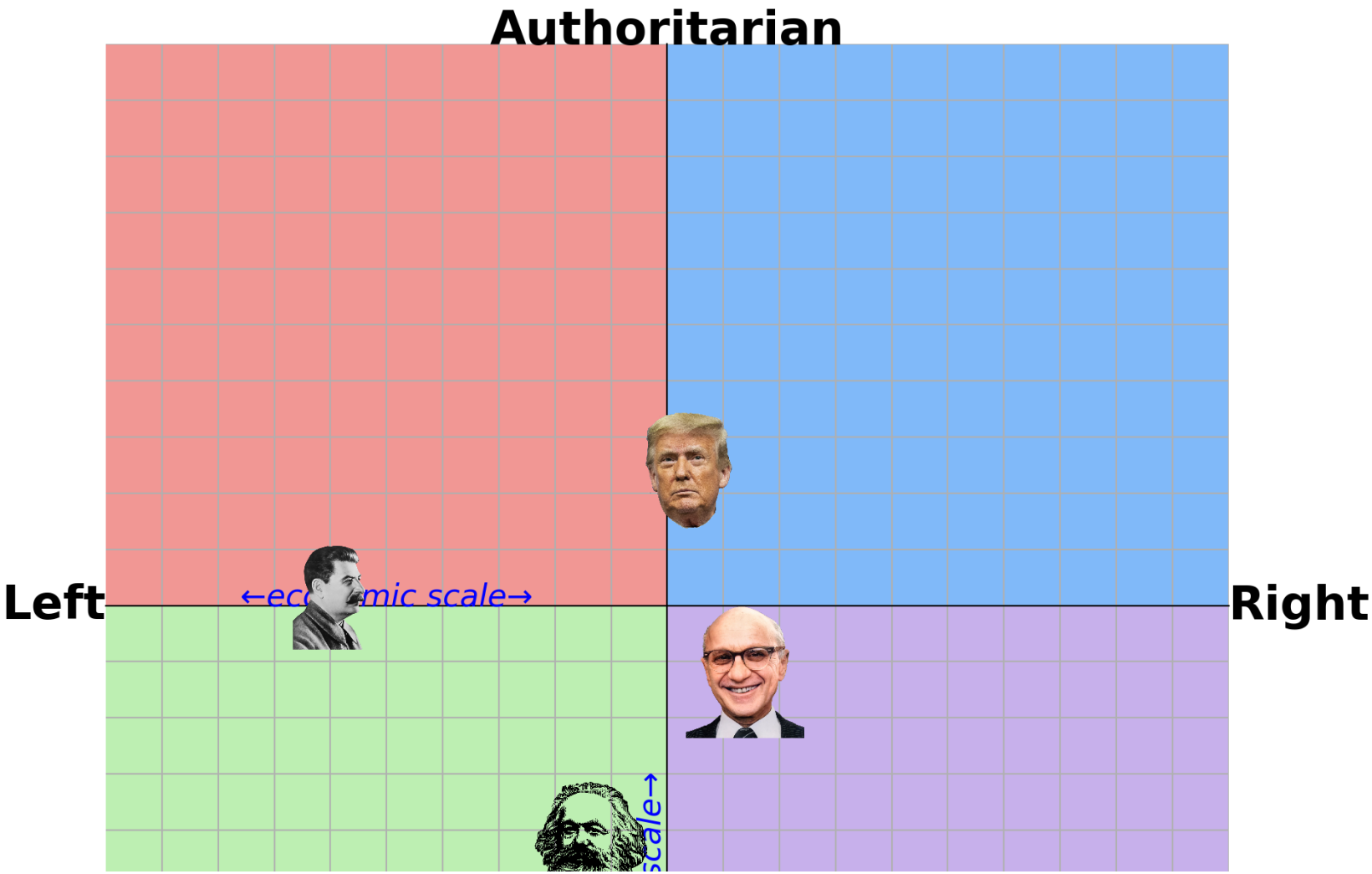


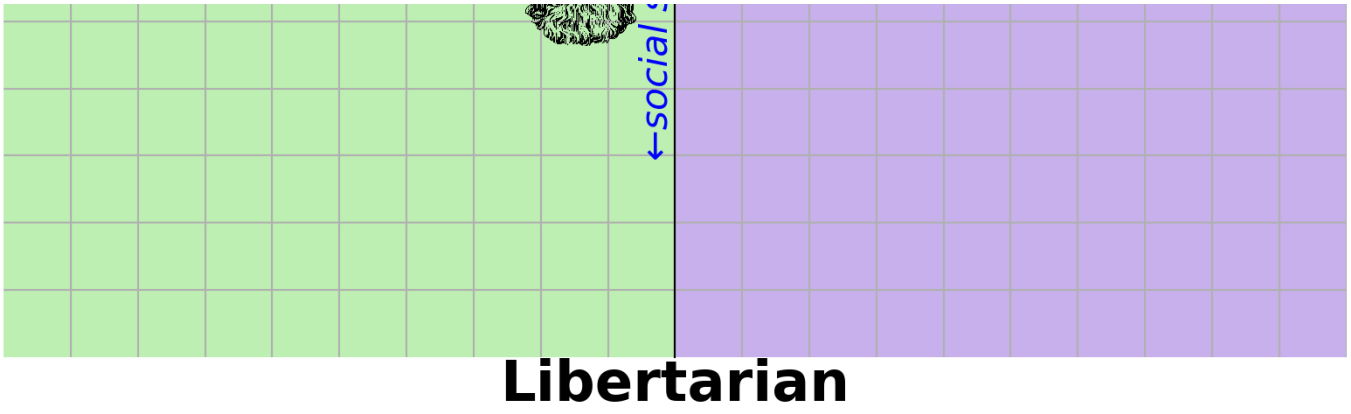
Visualisation of the Political Compass Test results.

Models include: Llama 3 8B Instruct (Meta), Qwen 2.5 72B Instruct (Qwen), Mixtral 8x7B Instruct (Mistral), Yi 34B Chat (Yi), and Grok-beta (xAI).

27

Output Code





Visualisation of the Political Compass Test results. Characters (Joseph Stalin, Karl Marx, Donald J Trump, and Milton Friedman) are persona from character.ai [33].

References

1. Feng S, Park CY, Liu Y, Tsvetkov Y (2023) From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers). Association for Computational Linguistics, Toronto, Canada, pp 11737–11762
2. Tao Y, Viberg O, Baker RS, Kizilcec RF (2023) Auditing and mitigating cultural bias in llms. arXiv preprint arXiv:231114096
3. Rutinowski J, Franke S, Endendyk J, Dormuth I, Roidl M, Pauly M (2024) The self-perception and political biases of ChatGPT. Human Behavior and Emerging Technologies 2024(1):7115633
4. Yakura H, Lopez-Lopez E, Brinkmann L, Serna I, Gupta P, Rahwan I (2024) Empirical evidence of large language model's influence on human spoken communication
5. Geng M, Chen C, Wu Y, Chen D, Wan Y, Zhou P (2024) The impact of large language models in academia: From writing to speaking
6. Anderson BR, Shah JH, Kreminski M (2024) Homogenization effects of large language models on human creative ideation. In: Creativity and cognition. ACM, pp 413–425

7. Jakesch M, Bhat A, Buschek D, Zalmanson L, Naaman M (2023) Co-writing with opinionated language models affects users' views. In: Proceedings of the 2023 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA
8. Li C, Su X, Han H, Xue C, Zheng C, Fan C (2023) Quantifying the impact of large language models on collective opinion dynamics. arXiv preprint arXiv:230803313
9. Williams-Ceci S, Jakesch M, Bhat A, et al Bias in AI autocomplete suggestions leads to attitude shift on societal issues
10. Jiang W (2000) The relationship between culture and language. ELT journal 54(4):328–334
11. Poole-Dayan E, Roy D, Kabbara J (2024) LLM targeted underperformance disproportionately impacts vulnerable users
12. Cunningham J, Blodgett SL, Madaio M, Daumé Iii H, Harrington C, Wallach H (2024) Understanding the impacts of language technologies' performance disparities on African American language speakers. In: Ku L-W, Martins A, Srikumar V (eds) Findings of the association for computational linguistics: ACL 2024. Association for Computational Linguistics, Bangkok, Thailand, pp 12826–12833
13. Salinas A, Shah P, Huang Y, McCormack R, Morstatter F (2023) The unequal opportunities of large language models: Examining demographic biases in job recommendations by ChatGPT and LLaMA. In: Proceedings of the 3rd ACM conference on equity and access in algorithms, mechanisms, and optimization.

Association for Computing Machinery, New York, NY, USA
14. Hadar-Shoval D, Asraf K, Mizrachi Y, Haber Y, Elyoseph Z (2023) The invisible embedded “values” within large language models: Implications for mental health use
15. Liu Y, Feng S, Han X, et al (2024) P^3 Sum: Preserving author’s perspective in news summarization with diffusion language models. In: Duh K, Gomez H, Bethard S (eds) Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers). Association for Computational Linguistics, Mexico City, Mexico, pp 2154–2173
16. Vaswani A (2017) Attention is all you need. Advances in Neural Information Processing Systems
17. Liu Y, Cao J, Liu C, Ding K, Jin L (2024) Datasets for large language models: A comprehensive survey. arXiv preprint arXiv:240218041
18. Stollnitz B (2023) The transformer architecture of GPT models. Bea Stollnitz Blog
19. Röttger P, Hofmann V, Pyatkin V, et al (2024) Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In: Ku L-W, Martins A, Srikumar V (eds) Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers). Association for Computational Linguistics, Bangkok, Thailand, pp 15295–15311
20. Wright D, Arora A, Borenstein N, Yadav S, Belongie S, Augenstein I (2024) LLM

tropes: Revealing fine-grained values and opinions in large language models. In: Al-Onaizan Y, Bansal M, Chen Y-N (eds) Findings of the association for computational linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp 17085–17112

21. Motoki F, Pinho Neto V, Rodrigues V (2024) More human than human: Measuring ChatGPT political bias. Public Choice 198(1):3–23
22. Agiza A, Mostagir M, Reda S (2024) PoliTune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society. pp 2–12
23. Kovač G, Sawayama M, Portelas R, Colas C, Dominey PF, Oudeyer P-Y (2023) Large language models as superpositions of cultural perspectives. arXiv preprint arXiv:230707870
24. Wang H, Zhao S, Qiang Z, Xi N, Qin B, Liu T (2024) Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. arXiv preprint arXiv:240201349
25. Khatun A, Brown DG (2024) A study on large language models' limitations in multiple-choice question answering. arXiv preprint arXiv:240107955
26. Tsvilodub P, Wang H, Grosch S, Franke M (2024) Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods. arXiv preprint arXiv:240300998
27. Wang X, Ma B, Hu C, et al (2024) "My answer is C": First-token probabilities do not

match text answers in instruction-tuned language models. In: Ku L-W, Martins A, Srikumar V (eds) Findings of the association for computational linguistics: ACL 2024. Association for Computational Linguistics, Bangkok, Thailand, pp 7407–7416

28. Balepur N, Ravichander A, Rudinger R (2024) Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In: Ku L-W, Martins A, Srikumar V (eds) Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers). Association for Computational Linguistics, Bangkok, Thailand, pp 10308–10330
29. Ranaldi L, Zanzotto F (2024) HANS, are you clever? Clever hans effect analysis of neural systems. In: Bollegala D, Shwartz V (eds) Proceedings of the 13th joint conference on lexical and computational semantics (*SEM 2024). Association for Computational Linguistics, Mexico City, Mexico, pp 314–325
30. Abercrombie G, Curry AC, Dinkar T, Rieser V, Talat Z (2023) Mirages: On anthropomorphism in dialogue systems. arXiv preprint arXiv:230509800
31. Zhao W, Ren X, Hessel J, Cardie C, Choi Y, Deng Y (2024) WildChat: 1M chatGPT interaction logs in the wild. In: The twelfth international conference on learning representations
32. Bang Y, Chen D, Lee N, Fung P (2024) Measuring political bias in large language models: What is said and how it is said. In: Ku L-W, Martins A, Srikumar V (eds) Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers). Association for Computational Linguistics,