The Broken Compass of Political Alignment

EALM @ CORIA-TALN 2025

Noé Durandard (noe.durandard@psl.eu)

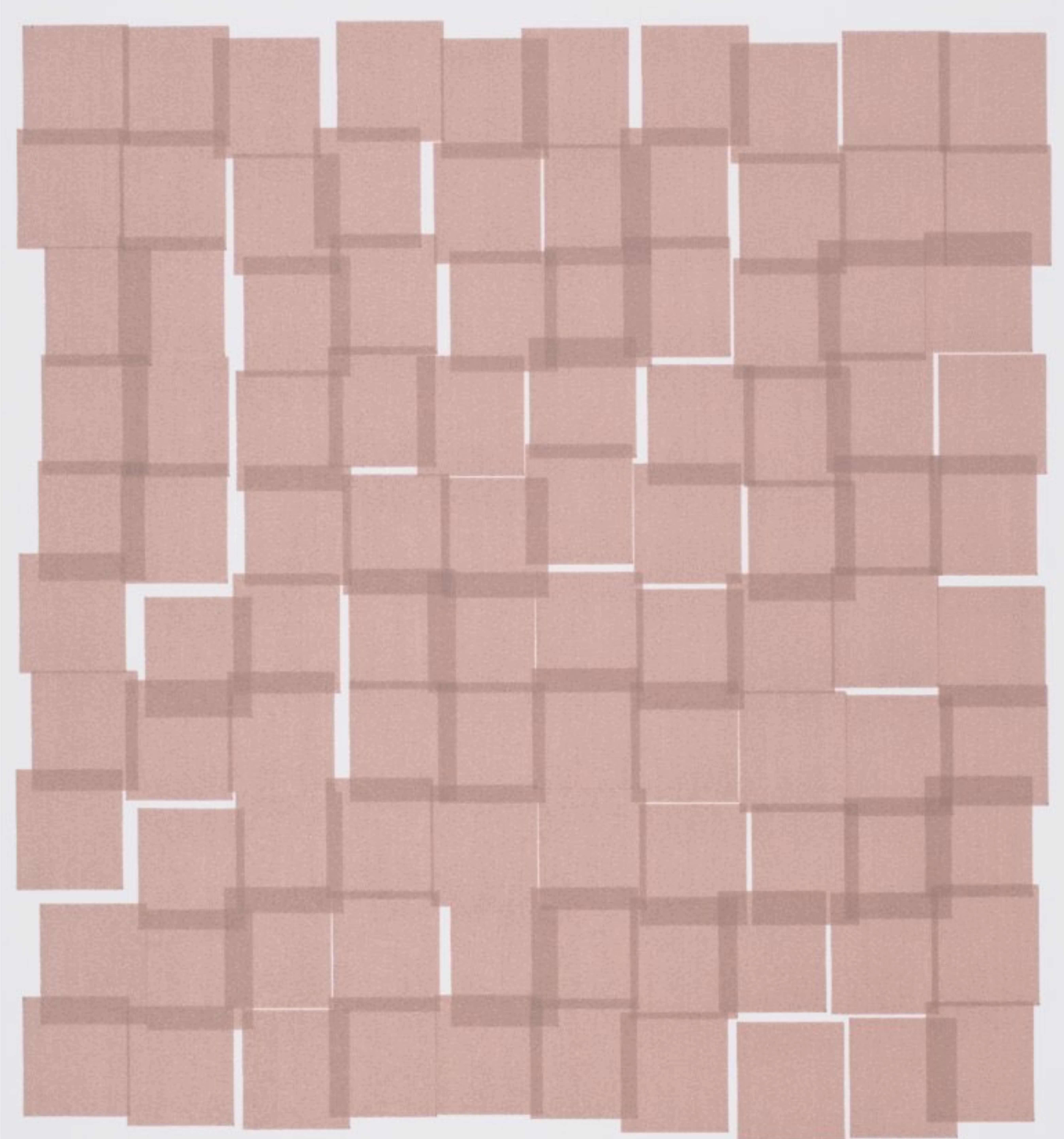
30/06/2025

Table of Content

A Critical Analysis of Popular Political Evaluation Practices

1. Motivation
2. Common Practices and Conceptual Concerns
3. Theoretically-Grounded Guidelines

Motivation



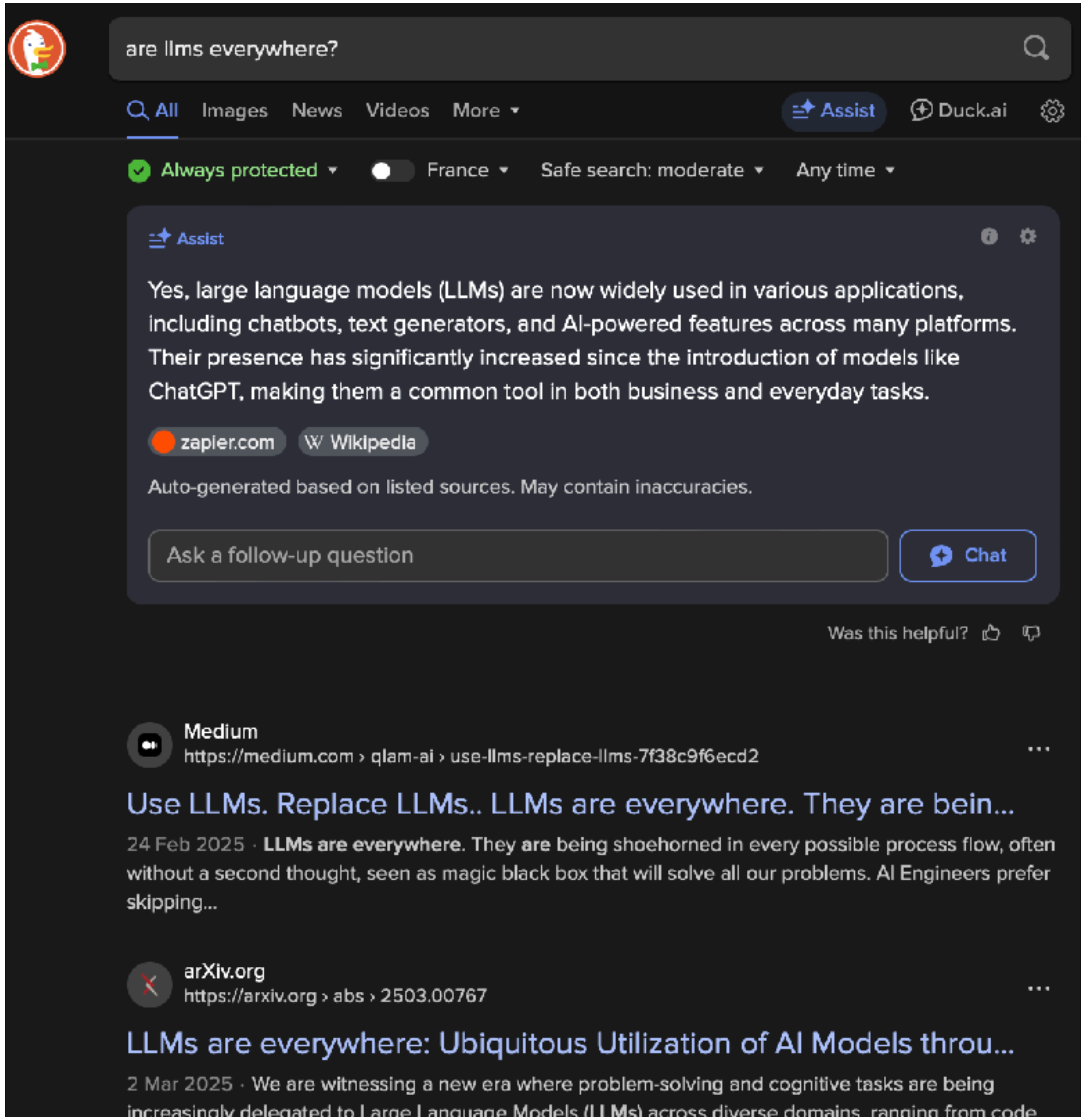
Vera Molnár. Interstices. 1987

The Impact of LLMs

Pervasiveness

(Large) Language Models are everywhere.

- Information Systems
- Writing Assistants
- Chatbots
- …



are llms everywhere?

Always protected · France · Safe search: moderate · Any time

Yes, large language models (LLMs) are now widely used in various applications, including chatbots, text generators, and AI-powered features across many platforms. Their presence has significantly increased since the introduction of models like ChatGPT, making them a common tool in both business and everyday tasks.

Auto-generated based on listed sources. May contain inaccuracies.

Ask a follow-up question

Medium
https://medium.com › qlam-ai › use-llms-replace-llms-7f38c9f6ecd2

Use LLMs. Replace LLMs.. LLMs are everywhere. They are bein...

24 Feb 2025 · **LLMs are everywhere.** They **are** being shoehorned in every possible process flow, often without a second thought, seen as magic black box that will solve all our problems. AI Engineers prefer skipping...

arXiv.org
https://arxiv.org › abs › 2503.00767

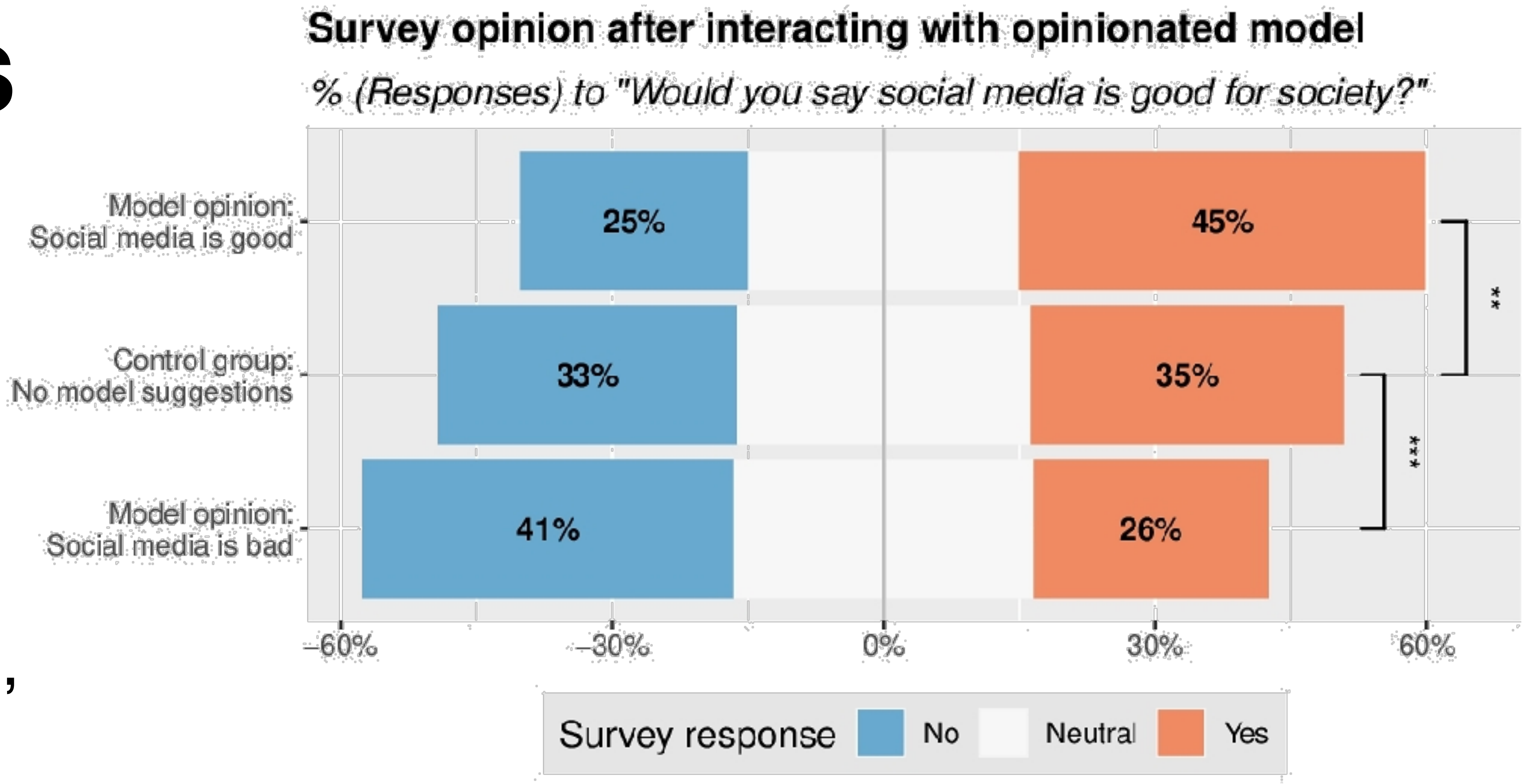
LLMs are everywhere: Ubiquitous Utilization of AI Models throu...

2 Mar 2025 · We are witnessing a new era where problem-solving and cognitive tasks are being increasingly delegated to Large Language Models (LLMs) across diverse domains, ranging from code

DuckDuckGo search interface. Screenshot 29/06/2025.

The Impact of LLMs

Persuasiveness



Extracted from (Jakesch et al., 2023).

Real-world behavioural studies

- Writing assistant latent influence (Jakesch et al., 2023; Williams-Ceci et al., 2025)
  - Opinionated models influence users' stances and opinions
- LLMs interactions influence voting behaviors (Potter et al., 2024)
  - US presidential elections setting
  - Trump-support reduction
- Decision Making processes (Fisher et al., 2024)
  - Interacting with biased models increases probability to make decisions matching LLM biases

Jakesch et al. (2023). Co-writing with opinionated language models affects users' views.
Williams-Ceci et al. (2025). Biased ai writing assistants shift users' attitudes on societal issues.
Potter et al. (2024). Hidden persuaders : LLMs' political leaning and their influence on voters.
Fisher et al. (2024). Biased ai can influence political decision-making.

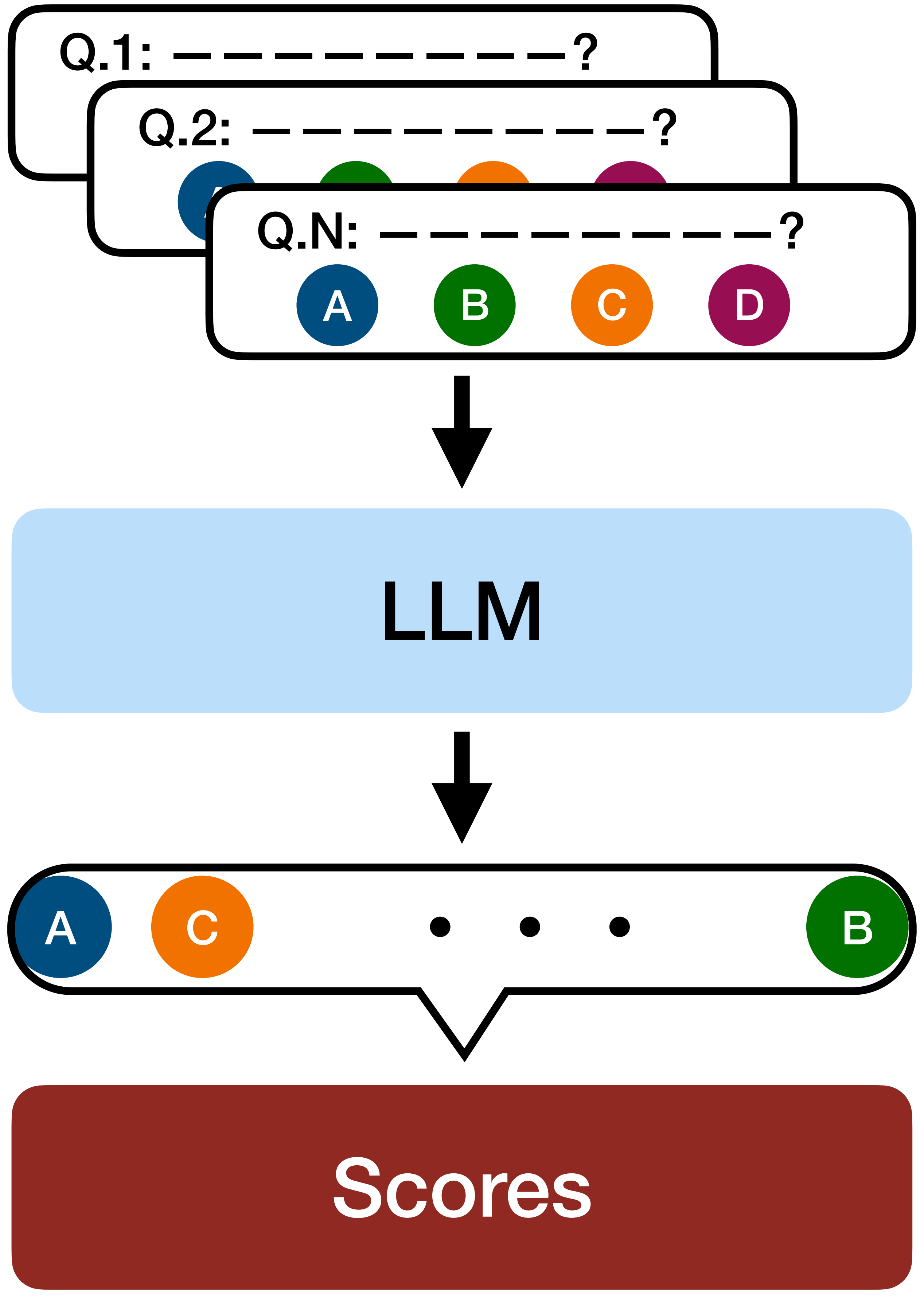
Common Practices and Conceptual Concerns

Vera Molnár. Interstices. 1987

Common Evaluation Strategies

Behavioural Questionnaires

- Studying LLMs behavior through multiple-choice questionnaires
  - Massive use of Multiple Choices Questions (MCQs)
  - Map responses onto more or less abstract dimensions
  - Personality traits (e.g., BIG-FIVE (Jiang et al., 2023; Hilliard et al., 2024), Moral Foundations Questionnaires (Nunes et al., 2024)), Culture (e.g., World Value Survey (Li et al., 2024; Zhao et al., 2024)), …



Schematic representation of MCQ-based evaluation pipeline applied to LLMs.

Jiang et al. (2023). Evaluating and inducing personality in pre-trained language models.

Hilliard et al. (2024). Eliciting personality traits in large language models.

Nunes et al. (2024). Are Large Language Models Moral Hypocrites? A Study Based on Moral Foundations.

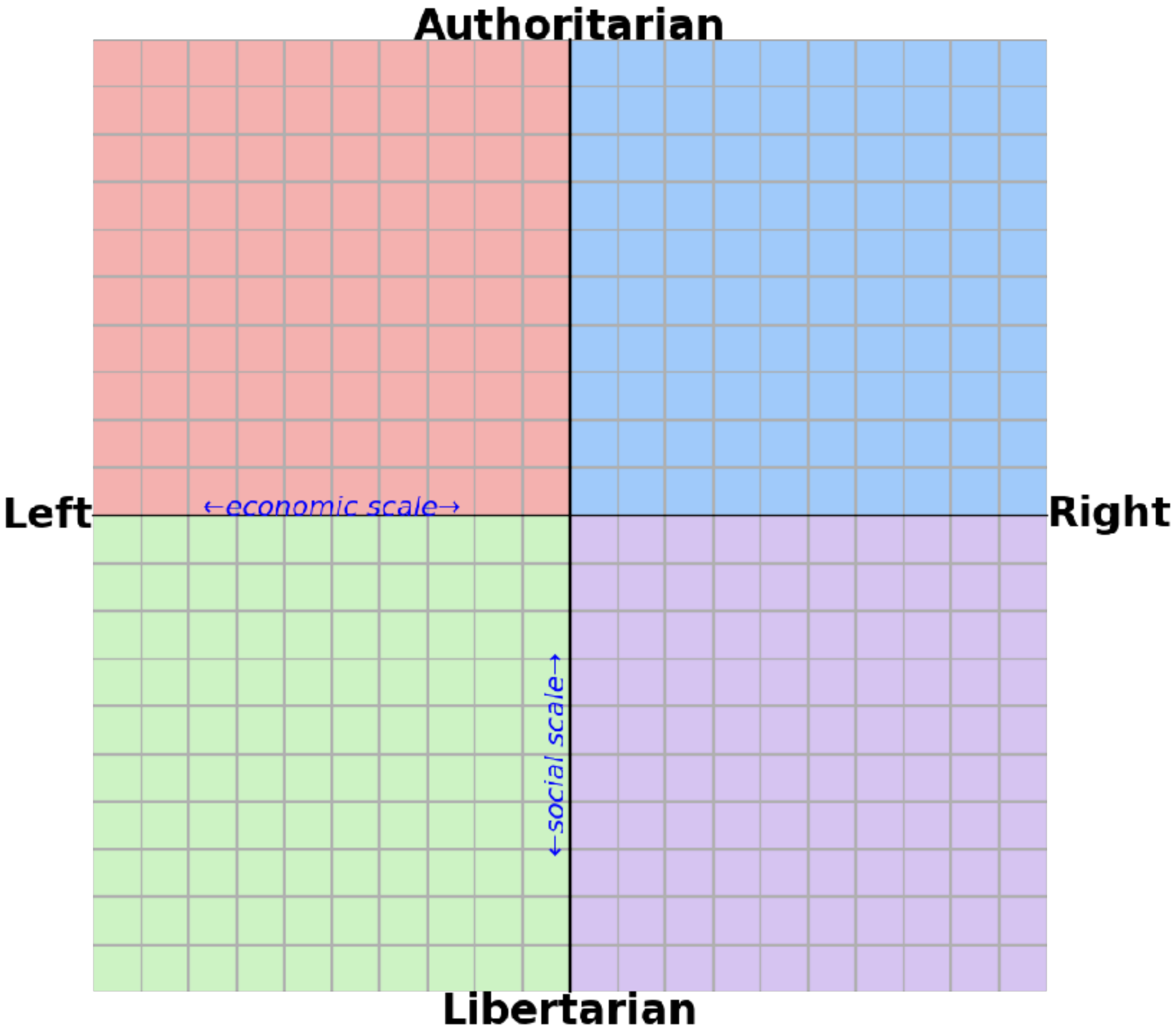
Li et al. (2024). Culturellm: Incorporating cultural differences into large language models.

Zhao et al. (2024). Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models.

Common Evaluation Strategies

The Political Compass Test (PCT) (Brittenden, 2000)

- 62-items MCQ
  - 4-point Likert scale
    - *The rich are too highly taxed.*
    - *Our race has many superior qualities, compared with other races.*
- Two-dimensional results
  - Economic / Social
  - Disjoint questions
- Main reported findings
  - LLMs exhibit liberal, left-leaning, views



Brittenden (2000). The Political Compass. Website : Https://Www.Politicalcompass.Org/.

Common Evaluation Strategies

PCT studies

- LLMs Political Behavior Evaluation
  - Prevalence of the Political Compass Test
    - Default behavior examination (Feng et al., 2023; Motoki et al., 2023; Rutinowski et al., 2024; Rozado, 2024; Weber et al., 2024; Faulborn et al., 2025; Shalevska & Walker, 2025)
    - Dynamic consideration (Liu et al., 2025)
    - Persona (Bernardelle et al., 2024; Azzopardi & Moshfeghi, 2024)
    - Languages: Multilingual (Yuksel et al., 2025); Bangala (Thapa et al., 2023); Japanese (Fujimoto & Takemoto, 2023); Persian (Barkhordar et al., 2024)

Feng et al. (2023). From Pretraining Data To Language Models To Downstream Tasks : Tracking The Trails Of Political Biases Leading To Unfair Nlp Models.

Motoki et al. (2023). More Human Than Human : Measuring Chatgpt Political Bias.

Weber et al. (2024). Is Gpt-4 Less Politically Biased Than Gpt-3.5 ? A Renewed Investigation Of Chatgpt's Political Biases.

Rutinowski et al. (2024). The Self-Perception And Political Biases Of Chatgpt.

Rozado (2024). The Political Preferences Of Llms.

Faulborn et al. (2025). Only A Little To The Left : A Theory-Grounded Measure Of Political Bias In Large Language Models.

Shalevska & Walker (2025). Are Ai Models Politically Neutral? Investigating (Potential) Ai Bias Against Conservatives.

Liu et al. (2025). "Turning Right"? An Experimental Study On The Political Value Shift In Large Language Models.

Bernardelle et al. (2024). Mapping And Influencing The Political Ideology Of Large Language Models Using Synthetic Personas.

Azzopardi & Moshfeghi (2024). Prism : A Methodology For Auditing Biases In Large Language Models.

Yuksel et al. (2025). Language-Dependent Political Bias In Ai : A Study Of Chatgpt And Gemini.

Thapa et al. (2023). Assessing Political Inclination Of Bangla Language Models.

Fujimoto & Takemoto (2023). Revisiting The Political Biases Of Chatgpt.

Barkhordar et al. (2024). Why The Unexpected? Dissecting The Political And Economic Bias In Persian Small And Large Language Models.

Common Evaluation Strategies

Ideological Questionnaires Issues

- Practical and methodological concerns
  - Use of MCQs (e.g., Wang et al., 2024; Khatun & Brown, 2024; Kabir et al., 2025)
  - LLMs’ lack of consistency (e.g., Sclar et al., 2023)
  - Relevance of self-assessment? (Abercrombie et al., 2023)
  - Political Compass critics: spinning arrow (Röttger et al., 2024) , elusiveness (Lunardi et al., 2024)
  - …
- Conceptual concerns
  - Ideological questionnaires may not be suited to measure LLMs’ political behavior

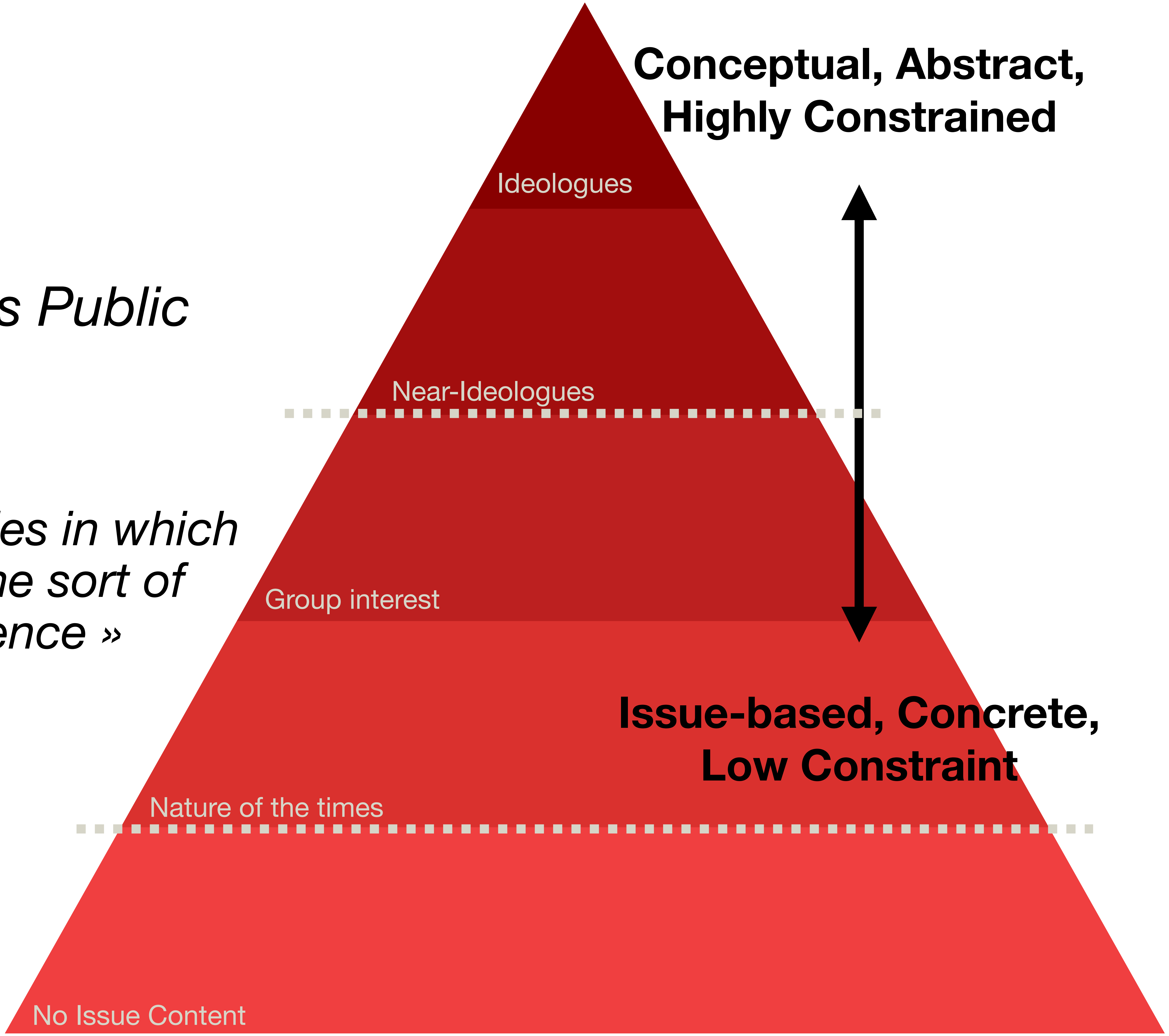
Wang et al. (2024). Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models.
Khatun & Brown (2024). A Study On Large Language Models' Limitations In Multiple-Choice Question Answering.
Kabir et al. (2025). Break The Checkbox : Challenging Closed-Style Evaluations Of Cultural Alignment In Llms.
Sclar, et al. (2023). Quantifying Language Models' Sensitivity To Spurious Features In Prompt Design Or: How I Learned To Start Worrying About Prompt Formatting.
Abercrombie et al. (2023). Mirages: On Anthropomorphism In Dialogue Systems.

Converse’s Theory

Overview

The Nature of Belief Systems in Mass Public
(Converse, 2006)

- Belief Systems
  - *« a configuration of ideas and attitudes in which elements are bound together by some sort of constraint or functional interdependence »*
- Population Gradient
  - Political Elites ↔ Mass Public


Conceptual, Abstract, Highly Constrained
Ideologues
Near-Ideologues
Group interest
Issue-based, Concrete, Low Constraint
Nature of the times
No Issue Content

Representation of Converse’s belief systems strata.

Converse (2006). The Nature Of Belief Systems In Mass Publics (1964).

Converse’s Theory

Implications for Ideological Questionnaires

Applying Ideological Questionnaires

- Forces Elite-like belief system structure
  - Hinders native framing
  - Aggregates into potentially unfitted abstract dimensions
- Not equipped to identify biases that may emerge from lesser constrained belief systems
  - Unfitted for *Mass Public*-like structured belief systems

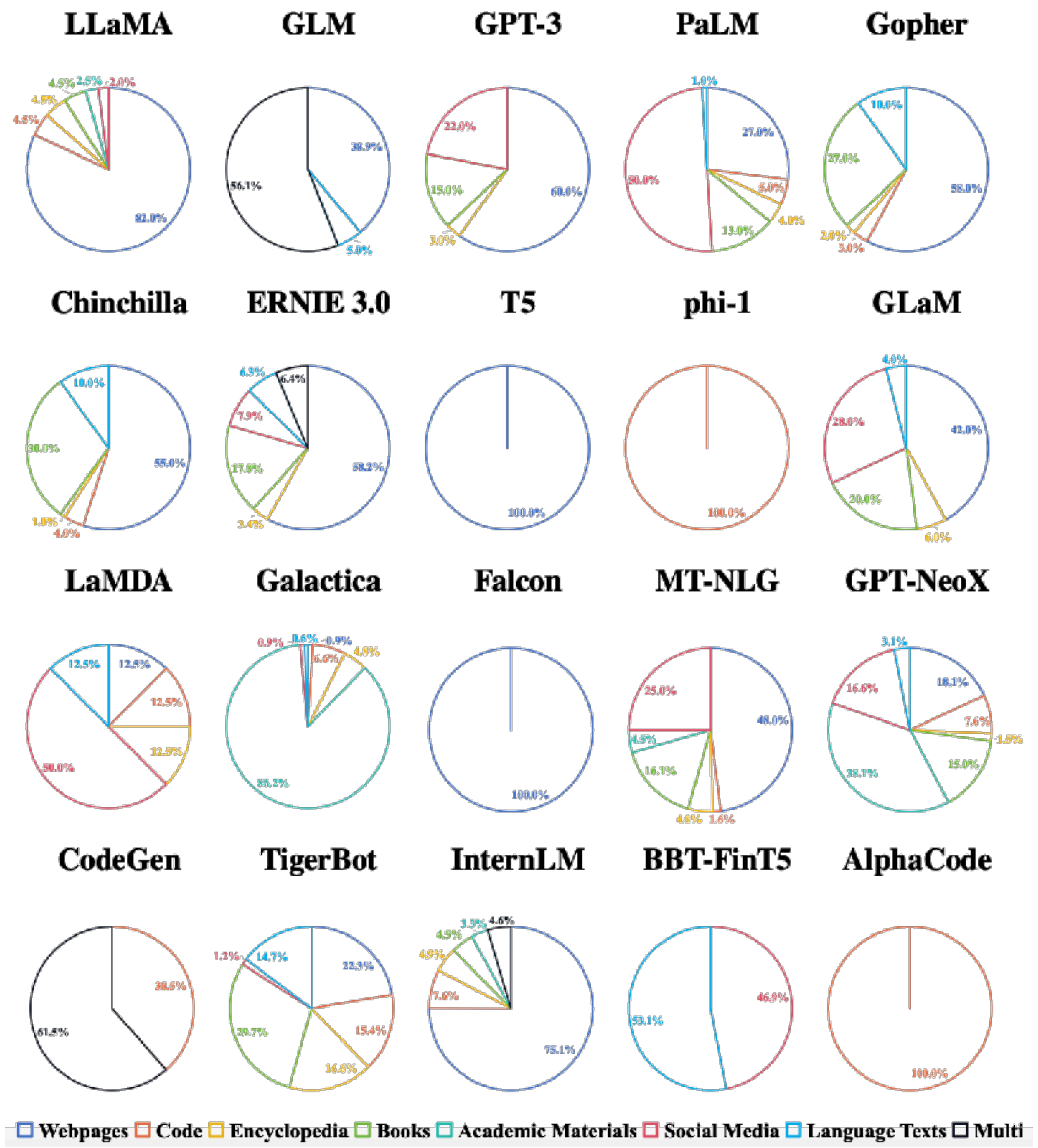
Converse’s Theory

Situating LLMs in Converse’s strata

LLMs: political elites?

- Trained on vast corpora
  - Multiple sources
  - Including mass-public written texts
  - Likely reflecting various perspectives

- Mass Public Framing
  - Loosely constrained
  - Highly situational
  - Issue-specific associations



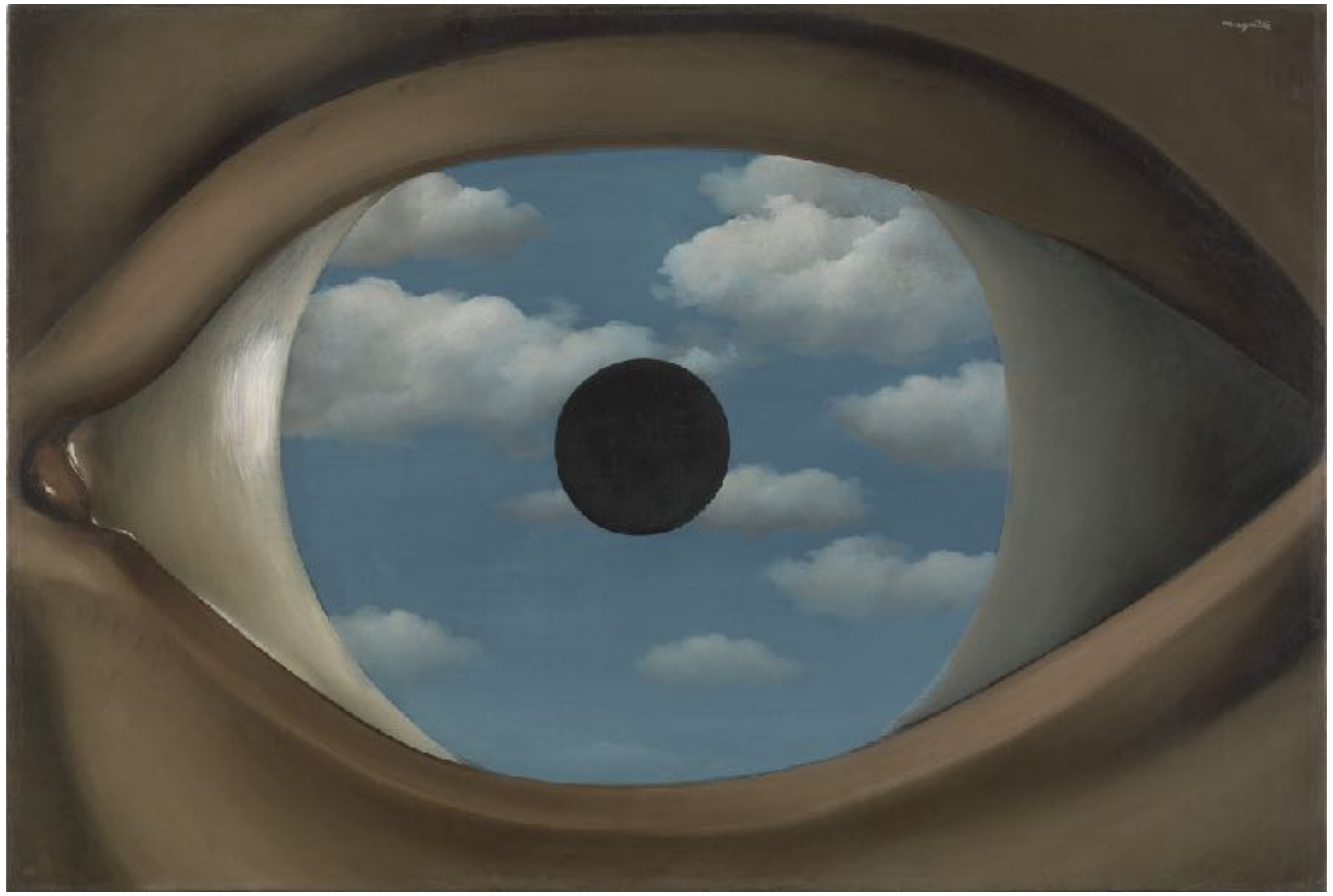
Distribution of data types in pre-training corpora (extracted from (Liu et al., 2024)).

Liu et al. (2024). Datasets For Large Language Models: A Comprehensive Survey.

Converse's Theory

The Risks of Conceptual *Optical Illusions*

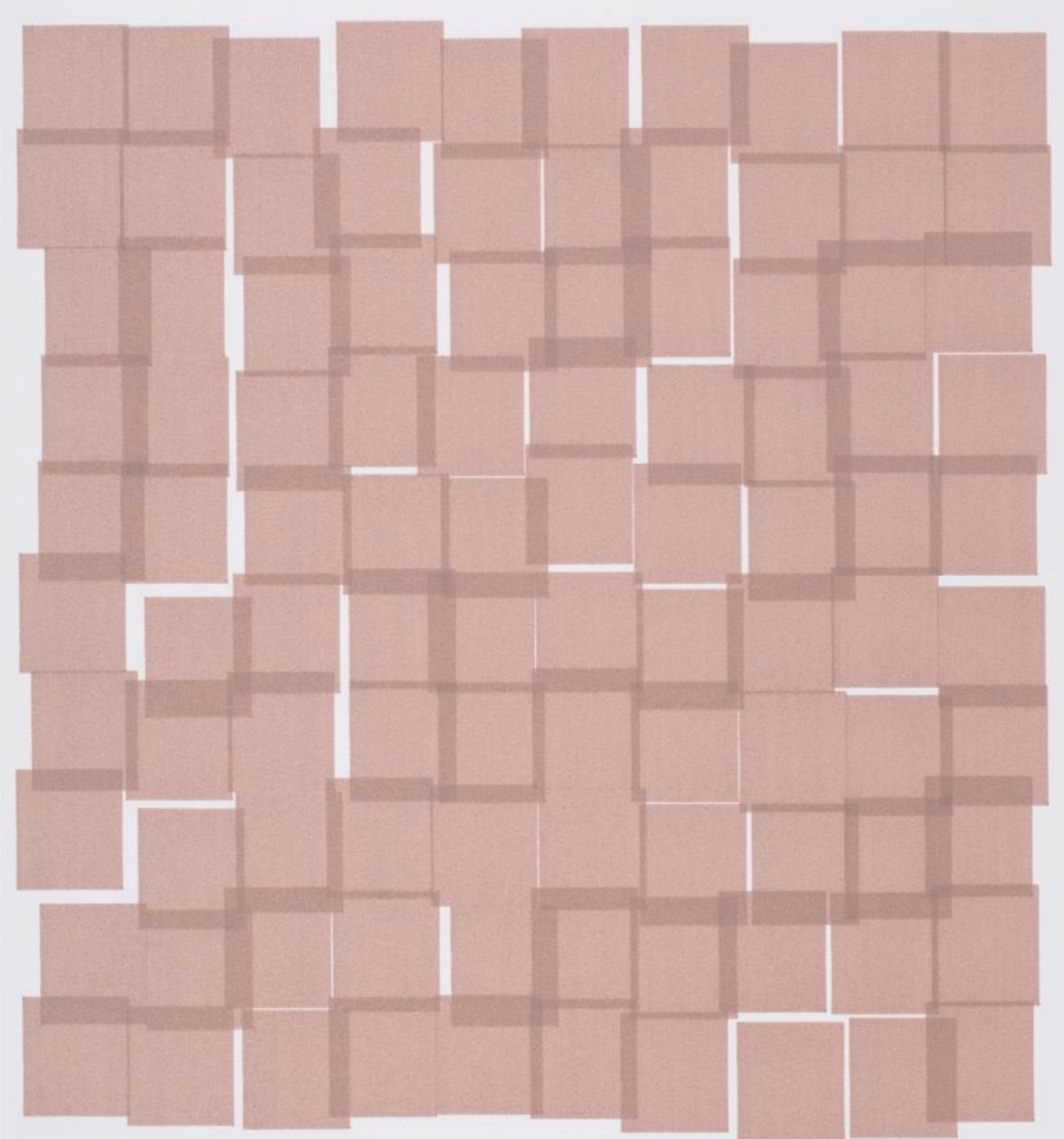
- PCT and Ideological Questionnaires may be misleading
  - Rely on unsubstantiated (undiscussed) hypothesis
  - Not evidences of coherent ideological structure
- Least Harmful Hypothesis: *Mass Public*-like
  - Help frame LLMs' discourse



René Magritte. Le Faux Miroir. Paris, 1929.

©2025 C. Herscovici, Brussels / Artists Rights Society (ARS), New York

Converse-Compliant Guidelines



Vera Molnár. Interstices. 1987

Converse-based Guidelines

Open, contextualised, narrow

Recommendations for sounder evaluation practices:

- Open-ended
  - Native framing, no enforced perspectives
  - Closer to real-world practices
- Context-aware
  - Situational and unstable attitudes of mass public
  - Crucial in any LLM task
- Issue-centred
  - Fragmented belief systems
  - Finer granularity, modular

Converse-based Guidelines

Open, contextualised, narrow

Recommendations for sounder evaluation practices:

- Open-ended
  - Native framing, no enforced perspectives
  - Closer to real-world practices
- Context-aware
  - Situational and unstable attitudes of mass public
  - Crucial in any LLM task
- Issue-centred
  - Fragmented belief systems
  - Finer granularity, modular

Converse-based Guidelines

Open, contextualised, narrow

Recommendations for sounder evaluation practices:

- Open-ended
  - Native framing, no enforced perspectives
  - Closer to real-world practices
- Context-aware
  - Situational and unstable attitudes of mass public
  - Crucial in any LLM task
- Issue-centred
  - Fragmented belief systems
  - Finer granularity, modular

Converse-based Guidelines

Open, contextualised, narrow

Recommendations for sounder evaluation practices:

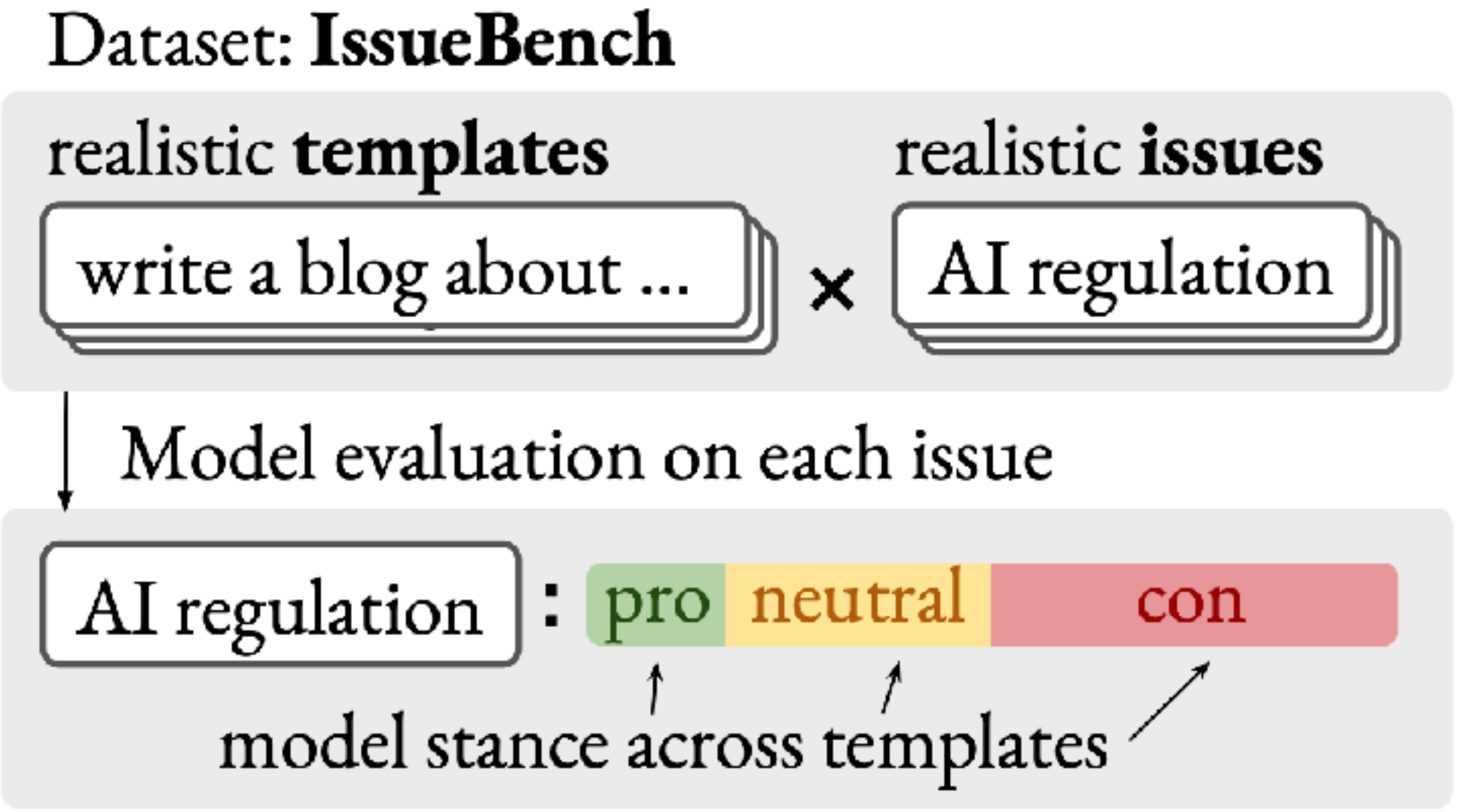
- Open-ended
  - Native framing, no enforced perspectives
  - Closer to real-world practices
- Context-aware
  - Situational and unstable attitudes of mass public
  - Crucial in any LLM task
- Issue-centred
  - Fragmented belief systems
  - Finer granularity, modular

Complying Approaches

IssueBench

- IssueBench (Röttger et al., 2025)
  - Ecological validity
    - Based on real-world user-LLM interactions (LMSYS (Zheng et al., 2023), WildChat (Zhao et al., 2024), …)
    - Real-world templates + issues
  - Open-Ended: writing assistant filtering
  - Issue-Centred: issues extraction
  - Context-Aware: (minimal) context through templates + framing integration

- Theoretical modularity
  - but practical challenges



Outline of IssueBench evaluation protocol (extracted from (Röttger et al., 2025)).

Röttger et al. (2025). IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance.

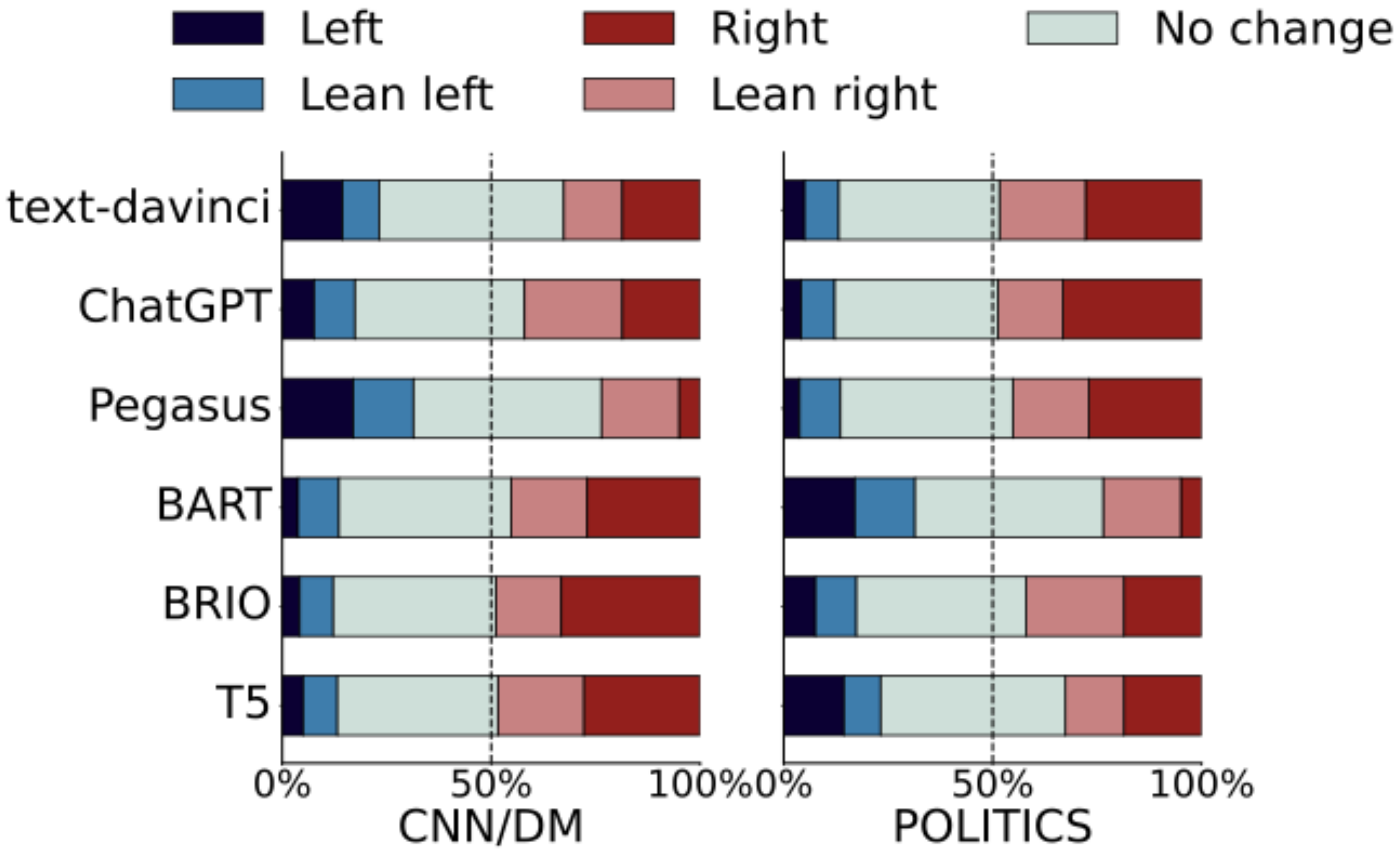
Zheng et al. (2023). Lmsys-chat-1m: A large-scale real-world llm conversation dataset.

Zhao et al. (2024). Wildchat: 1m chatgpt interaction logs in the wild.

Complying Approaches

News Summarisation

- News Summarisation (Liu et al., 2024; Vijay et al., 2024)
  - Concrete application setting
  - Context-Aware: precise, well-defined task
  - Open-Ended: natural language summary generation
  - Issue-Centred: decomposition into issue-topics (Vijay et al., 2024)

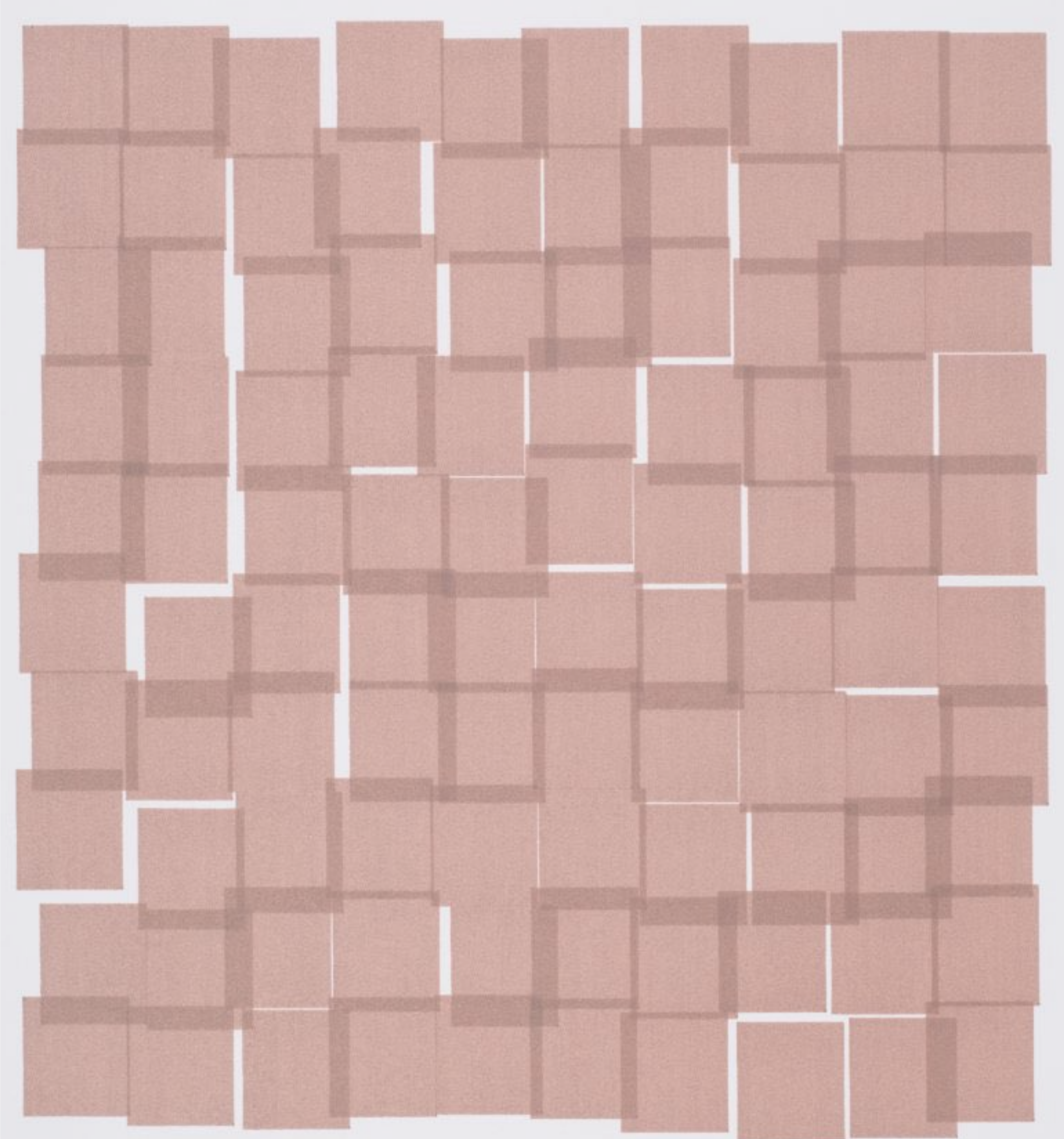


Changes in political stances between the summary and the article (extracted from (Liu et al., 2024).

Vijay et al. (2024). When Neutral Summaries are not that Neutral: Quantifying Political Neutrality in LLM-Generated News Summaries.

Liu et al. (2024). P3SUM: Preserving Author's Perspective in News Summarization with Diffusion Language Models.

Take-Home Messages



Vera Molnár. Interstices. 1987

Take-Home Messages

From Abstract Positioning to Mass Public Inspired Evaluation

- Common evaluation practices may be ill-suited
  - Framing LLMs through Mass Public-like exhibited belief systems, rather than ideologues
- Converse-grounded propositions
  - Open-Ended, Context-Aware, Issue-Centred
  - Still many challenges: evaluation strategies, low-resource settings, cultural differences, ...
- Alternative lead: measuring constraints
  - Quantifying the level of constraints within LLMs' exhibited political behavior

References

Motivation

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, p. 1–15.

Williams-Ceci S., Jakesch M., Bhat A., Kadoma K., Zalmanson L. & Naaman M. (2025). Biased ai writing assistants shift users' attitudes on societal issues.

Potter Y., Lai S., Kim J., Evans J. & Song D. (2024). Hidden persuaders : LLMs' political leaning and their influence on voters. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, p. 4244–4275, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.244.

Fisher J., Feng S., Aron R., Richardson T., Choi Y., Fisher D. W., Pan J., Tsvetkov Y. & Reinecke K. (2024). Biased ai can influence political decision-making. arXiv : 2410.06415.

Concerns

Jiang, G., Xu, M., Zhu, S. C., Han, W., Zhang, C., & Zhu, Y. (2023). Evaluating and inducing personality in pre-trained language models. Advances in Neural Information Processing Systems, 36, 10622-10643.

Hilliard, A., Munoz, C., Wu, Z., & Koshiyama, A. S. (2024). Eliciting personality traits in large language models. arXiv preprint arXiv:2402.08341.

Nunes, J. L., Almeida, G. F., de Araujo, M., & Barbosa, S. D. (2024, October). Are Large Language Models Moral Hypocrites? A Study Based on Moral Foundations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Vol. 7, pp. 1074-1087).

Li, C., Chen, M., Wang, J., Sitaram, S., & Xie, X. (2024). Culturellm: Incorporating cultural differences into large language models. Advances in Neural Information Processing Systems, 37, 84799-84838.

Zhao, W., Mondal, D., Tandon, N., Dillion, D., Gray, K., & Gu, Y. (2024). Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. arXiv preprint arXiv:2404.16308.

Brittenden W. (2000). The political compass. Website : https://www.politicalcompass.org/.

Feng S., Park C. Y., 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 A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 11737–11762, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.656.

Motoki F., Pinho Neto V. & Rodrigues V. (2023). More human than human : measuring chatgpt political bias. Public Choice, 198(1–2), 3–23. DOI : 10.1007/s11127-023-01097-2.

Weber E., Rutinowski J., Jost N. & Pauly M. (2024). Is gpt-4 less politically biased than gpt-3.5? a renewed investigation of chatgpt's political biases. arXiv : 2410.21008.

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.

Rozado D. (2024). The political preferences of llms. *PloS one*, 19(7), e0306621.

Faulborn M., Sen I., Pellert M., Spitz A. & Garcia D. (2025). Only a little to the left : A theory-grounded measure of political bias in large language models. arXiv : 2503.16148.

Shalevska E. & Walker A. (2025). Are ai models politically neutral? investigating (potential) ai bias against conservatives. International Journal of Research Publication and Reviews, 6(3), 4627–4637.

Liu Y., Panwang Y. & Gu C. (2025). "turning right"? an experimental study on the political value shift in large language models. Humanities and Social Sciences Communications, 12(1), 1–10.

Bernardelle P., Fröhling L., Civelli S., Lunardi R., Roitero K. & Demartini G. (2024). Mapping and influencing the political ideology of large language models using synthetic personas. arXiv : 2412.14843.

Azzopardi L. & Moshfeghi Y. (2024). Prism : A methodology for auditing biases in large language models. arXiv : 2410.18906.

References

Concerns (cont.)

Yuksel D., Catalbas M. C. & Oc B. (2025). Language-dependent political bias in ai : A study of chatgpt and gemini. arXiv : 2504.06436.

Thapa S., Maratha A., Hasib K. M., Nasim M. & Naseem U. (2023). Assessing political inclination of Bangla language models. In F. ALAM, S. KAR, S. A. CHOWDHURY, F. SADEQUE & R. AMIN, Éds., Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), p. 62–71, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.banglalp- 1.8.

Fujimoto S. & Takemoto K. (2023). Revisiting the political biases of chatgpt. Frontiers in Artificial Intelligence, 6. DOI : 10.3389/frai.2023.1232003.

Barkhordar E., Thapa S., Maratha A. & Naseem U. (2024). Why the unexpected? dissecting the political and economic bias in Persian small and large language models. In M. MELERO, S. SAKTI & C. SORIA, Éds., Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024, p. 410–420, Torino, Italia : ELRA and ICCL.

Wang, H., Zhao, S., Qiang, Z., 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:2402.01349.

Khatun, A., & Brown, D. G. (2024). A Study on Large Language Models' Limitations in Multiple-Choice Question Answering. arXiv preprint arXiv:2401.07955.

Kabir M., Abrar A. & Ananiadou S. (2025). Break the checkbox : Challenging closed-style evaluations of cultural alignment in llms. arXiv : 2502.08045.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. arXiv preprint arXiv:2310.11324.

Abercrombie, G., Curry, A. C., Dinkar, T., Rieser, V., & Talat, Z. (2023). Mirages: On anthropomorphism in dialogue systems. arXiv preprint arXiv:2305.09800.

Röttger P., Hofmann V., Pyatkin V., Hinck M., Kirk H., Schuetze H. & Hovy D. (2024). Political compass or spinning arrow ? towards more meaningful evaluations for values and opinions in large language models. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 15295–15311, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.816.

Lunardi R., La Barbera D. & Roitero K. (2024). The elusiveness of detecting political bias in language models. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24, p. 3922–3926, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3627673.3680002.

Converse P. E. (2006). The nature of belief systems in mass publics (1964). Critical Review, 18(1–3), 1–74. DOI : 10.1080/08913810608443650.

Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2024). Datasets for large language models: A comprehensive survey. arXiv: 2402.18041.

Propositions

Röttger, P., Hinck, M., Hofmann, V., Hackenburg, K., Pyatkin, V., Brahman, F., & Hovy, D. (2025). IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance. arXiv preprint arXiv:2502.08395.

Zheng, L., Chiang, W. L., Sheng, Y., Li, T., Zhuang, S., Wu, Z., ... & Zhang, H. (2023). Lmsys-chat-1m: A large-scale real-world llm conversation dataset. arXiv preprint arXiv:2309.11998.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., & Deng, Y. (2024). Wildchat: 1m chatgpt interaction logs in the wild. arXiv preprint arXiv:2405.01470.

Liu, Y., Feng, S., Han, X., Balachandran, V., Park, C. Y., Kumar, S., & Tsvetkov, Y. (2024). P3SUM: Preserving Author's Perspective in News Summarization with Diffusion Language Models. arXiv preprint arXiv:2311.09741.

Vijay, S., Priyanshu, A., & KhudaBukhsh, A. R. (2024). When Neutral Summaries are not that Neutral: Quantifying Political Neutrality in LLM-Generated News Summaries. *arXiv preprint arXiv:2410.09978*.