Measuring Human-Like Bias in LLMs? A Critique of Human-Derived Bias Constructs in LLM Evaluation

Antonela Tommasel[1,3], **Markus Schedl**[1,2]
[1]Johannes Kepler University Linz, Austria
[2]Linz Institute of Technology, Linz, Austria
[3]ISISTAN, CONICET-UNICEN, Argentina
{antonela.tommasel, markus.schedl}@jku.at

Abstract

Researchers increasingly use human-derived bias constructs to study Large Language Models (LLMs), including social-cognitive constructs such as implicit bias and stereotype activation, and cognitive biases such as anchoring, framing effects, and confirmation bias. Such approaches offer alternatives to overt bias probes, particularly when direct questioning may obscure bias or when model behaviour appears normatively acceptable. However, adapting human bias constructs to LLMs introduces an inferential gap. Psychological instruments were developed to study human cognition and social behaviour, whereas LLM evaluations rely on probabilities, text completions, rankings, or simulated decisions. This paper critiques human-centered bias evaluation in LLMs. We show how this gap arises from mismatches pertaining to human-derived constructs, human-model differences, and evaluation contexts, which can blur distinct interpretations of model bias. We then introduce a framework providing an analytical lens for relating these elements to warranted interpretations, with attention to target constructs, operationalizations, scope of inference, and limits of human analogy.

1 Introduction

Bias evaluation in Large Language Models (LLMs) increasingly relies on constructs originally developed to explain human cognition and social behaviour (Sumita et al., 2025). Recent studies (e.g., (Hida et al., 2025)) test whether models exhibit stereotype-consistent associations, implicit-bias-like patterns, socially desirable responses, anchoring effects, framing sensitivity, confirmation bias, or demographic disparities. These approaches go beyond overt bias probes, where models are directly asked to express attitudes, judgments, or preferences, and beyond aggregate performance metrics that may obscure systematic variation across conditions. They are especially relevant when instruction tuning and safety training shape models toward normatively acceptable answers.

However, adapting human-derived bias constructs to LLMs introduces an *inferential gap*[1]. The gap arises from construct transfer: instruments such as implicit association tests, social desirability measures, anchoring and framing-effect tasks were developed to study human cognition, judgment, and behaviour (Greenwald and Banaji, 1995; Greenwald et al., 1998; Tversky and Kahneman, 1981, 1974). In LLM evaluation, these constructs are operationalized through different observables, including token probabilities, generated completions, rankings, refusals, answer changes across prompts, or simulated decisions (Sumita et al., 2025). Hence, *the relationship between the psychological construct being invoked and the model behaviour being measured is not straightforward.*

This paper presents a critique of human-centered bias evaluation in LLMs. We argue that the inferential gap matters because model-side evidence can support different kinds of claims. For example, if a model assigns higher probability to *man–career* than to *woman–career*, this may indicate an association in the model, but it does not by itself show that the model will generate biased recommendations, cause harm in a hiring-support setting, or exhibit human-like implicit bias. Although our focus is on bias evaluation, similar inferential concerns may arise in other uses of psychological constructs for LLM evaluation, such as personality, emotion, moral judgment, theory of mind or pragmatic reasoning. We focus on bias because human-derived constructs are frequently used in this area and because overextended interpretations can have practical consequences for how model behaviour, harm and human analogy are communicated. We introduce an analytical framework that

[1]Appendix A provides examples of how this gap appears in existing LLM bias-evaluation studies.

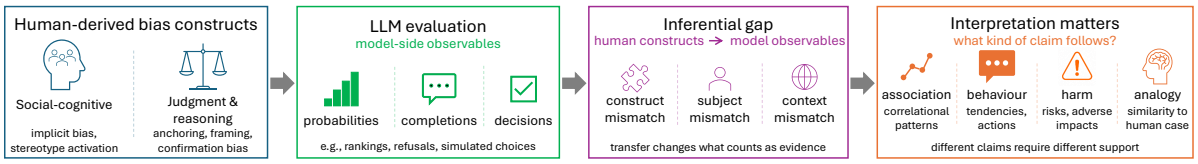

Figure 1: Conceptual overview. Human-derived bias constructs are operationalized through model-side observables in LLM evaluation, creating an inferential gap that motivates more precise interpretation of bias-related findings.

relates human-derived bias constructs, model-side observables, and warranted interpretations. *The goal is not to prescribe a new evaluation protocol, but to clarify what kinds of claims different evaluation designs can support and where stronger theoretical justification is required.* Figure 1 provides a conceptual overview.

2 From Evaluation of Human Bias Constructs to LLM Observables

To make the transfer problem of human bias constructs to LLMs more concrete, we distinguish two families of LLM bias probes, highlighting ways in which human-derived bias constructs are translated into model-side observables[2].

Social-cognitive bias. One line of work adapts concepts from social cognition, prejudice research, and fairness-oriented NLP to study whether LLMs reproduce socially meaningful associations or disparities. These evaluations are motivated by concerns about stereotyping, representational harms, and unequal treatment across demographic groups (Gallegos et al., 2024). Inspired by the Implicit Association Test (IAT) (Greenwald and Banaji, 1995; Greenwald et al., 1998), which infers relative associations from differences in response times to stimuli, recent work evaluates whether models assign higher probabilities to stereotype-consistent pairings, generate associations aligned with social stereotypes, or produce different outputs depending on demographic attributes (Bai et al., 2025; Zhao et al., 2025). Bias is operationalized through token likelihoods, embedding similarities, ranking preferences, controlled completions, or generation-based tasks involving demographic cues (Gallegos et al., 2024; Soundararajan and Delany, 2024; van Blerck et al., 2025). Related work studies socially desirable responses of LLMs, especially as instruction tuning, reinforcement learning from human feedback (RLHF), and safety alignment shape models toward normatively acceptable responses (Salecha et al., 2024; Ouyang et al., 2022; Bai et al., 2025).

[2]Appendix B provides a short example contrasting an IAT-inspired social-cognitive bias probe with an anchoring-style cognitive-bias probe.

Cognitive bias in judgment and reasoning. A second line of work investigates whether LLMs exhibit patterns analogous to human cognitive biases in judgment and reasoning. Drawing from behavioural economics and cognitive psychology, these evaluations test whether models reproduce deviations from normative reasoning under controlled task conditions (Echterhoff et al., 2024; Nguyen, 2024; Geva et al., 2025). Anchoring evaluations, for example, provide numerical anchors, prior estimates, or biased contextual information and measure whether outputs shift toward the anchor (Lou and Sun, 2026; Nguyen, 2024); framing evaluations test whether logically equivalent prompts produce different answers depending on wording or presentation (Tversky and Kahneman, 1981; Echterhoff et al., 2024). Other works study confirmation bias, overconfidence, or recency effects through reasoning tasks, decision scenarios, and prompt perturbations inspired by human experiments (Mina et al., 2025; Malberg et al., 2025).

Despite their different origins, social-cognitive and cognitive-bias evaluations share a common structure, a human-derived construct is translated into a model-side observable. These translations make psychological paradigms useful for probing LLMs, but they also create an inferential gap. *As LLMs differ from human participants in learning processes, mechanisms, and response modalities, the same construct label may refer to different kinds of evidence.* This raises questions about what kinds of interpretations model-side measurements can support (Cuskley et al., 2024; Mina et al., 2025).

3 The Inferential Gap in Human-Centered Bias Evaluation

This inferential gap is not simply a consequence of treating LLMs as another kind of human experimental subjects; this transfer changes what counts as evidence (Jacobs and Wallach, 2021; Gawronski et al., 2022). In psychology, bias constructs are embedded in theories that connect observable responses to assumptions about cognition, behaviour, and social context (Greenwald and Banaji, 1995; Tversky and Kahneman, 1981). In LLM evaluation, construct labels are attached to outputs pro-

duced by systems with different learning processes, mechanisms, and response modalities. The relation between construct, measurement, and interpretation therefore cannot be assumed to carry over unchanged (Cuskley et al., 2024).

Consider an LLM evaluation inspired by the IAT. In a human IAT (Greenwald and Banaji, 1995), response-time differences are used to infer associations between groups and attributes, such as woman/man and family/career. In an LLM adaptation (Gawronski et al., 2022), the evaluation instead compares likelihoods of pairings such as woman-family, woman-career, man-family, and man-career, or analyses gendered continuations in occupation-related prompts. This preserves the broad idea of indirect association measurement, but changes the subject, the form of response being measured, and the evidential basis of the task: reaction-time differences become likelihoods, rankings or completions from a model. Similarly, in an anchoring probe, a shift after an arbitrary cue may indicate prompt sensitivity, task behaviour, or analogy to human anchoring, depending on how the cue, output, and construct are connected.

Drawing on validity theory in psychology and social sciences (Cronbach and Meehl, 1955; Andrade, 2018; Cuskley et al., 2024), recent work framing generative AI evaluation as a social science measurement challenge (Wallach et al., 2025) and validity-threat analyses in empirical software engineering (Wohlin et al., 2012), we distinguish three sources of mismatch: *construct*, *subject*, and *evaluation-context*. These categories do not form a new validity theory; rather, they adapt familiar concerns about construct validity, external validity, and instrumentation to LLM bias evaluation.

Construct mismatch concerns whether a model-side measure captures the intended theoretical construct (Cronbach and Meehl, 1955; Jacobs and Wallach, 2021). Psychological bias constructs are embedded in theories connecting behaviour to assumptions about cognition, motivation, or reasoning. In the IAT example, the human construct is usually interpreted as an automatic association inferred indirectly from response-time differences (Greenwald and Banaji, 1995; Gawronski et al., 2022). In an LLM adaptation, the observable may be a likelihood difference between group-attribute pairings. This may show that some associations are stronger in the model than others (Bai et al., 2025). However, treating it as the same kind of implicit bias in humans requires explaining why a model probability should be interpreted as analogous to an indirect measure of automatic association.

Model-subject mismatch concerns the change from human participants to computational models. In a human IAT, the observed response pattern is produced by participants with embodied and social experience. LLMs are trained artifacts whose behaviour is shaped by pretraining data, fine-tuning, instruction following, RLHF, safety interventions, decoding strategies, and prompting (Ouyang et al., 2022; Holtzman et al., 2020). These differences do not make comparison impossible, but limit what can be inferred from behavioural similarity alone (Cuskley et al., 2024). Even if humans and LLMs produce similar association patterns, they might not have the same underlying reasoning or explanation.

Evaluation-context mismatch concerns how results depend on task setting, prompt design, and response modality. Human psychological tests and experiments are typically designed around specified constructs, administration conditions, response formats, scoring procedures, and validity evidence for the intended interpretations (Clark and Watson, 2019). LLM evaluations are controlled through prompts, system messages, examples, decoding, refusal policies, and conversational history, all of which can affect measured behaviour (Shi et al., 2024; Holtzman et al., 2020; Hida et al., 2025). In an LLM adaptation of the IAT, measured associations may vary with the prompt template, demographic labels, attribute words, examples, decoding settings, or whether the model completes text, ranks alternatives, or answers explicitly. The observed pattern therefore reflects both model behaviour and design choices in the evaluation setup, which may be hard to disentangle.

These mismatches do not imply that human-derived constructs are inappropriate for LLM evaluation. They can provide useful structure for designing probes and interpreting model behaviour patterns. The problem arises when model-side observables are treated as if they instantiate the human construct (Jacobs and Wallach, 2021). A likelihood difference, biased generation or answer shift may be meaningful evidence, but the claim it supports depends on the relation between the construct, observable behaviour, and evaluation context.

4 An Analytical Framework for Interpreting Bias Claims

The proposed analytical framework is inspired by validity theory in psychology and social sciences,

Interpretation	Claim	Typical model-side evidence	Mismatch(es)	Additional support needed
Distributional association	The model associates some groups, attributes, cues or outcomes more strongly than others.	Token probabilities, likelihood contrasts, embedding similarities, association scores.	Construct	Justification that the measured association is a meaningful proxy for the intended construct.
Observable behaviour	The model produces biased outputs under specific evaluation conditions.	Generated text, rankings, classifications, recommendations, answer shifts across prompts.	Subject, evaluation-context	Evidence that the behaviour is robust across relevant prompts, models, decoding settings or task variants.
Normative harm	The model behaviour may reinforce inequality, degrade decisions or cause harm in context.	Biased outputs connected to a use setting, affected users, groups or decisions.	Evaluation-context	An account of the deployment context, affected groups and plausible consequences.
Psychological analogy	The behaviour is interpreted as meaningfully similar to a human cognitive or social-cognitive bias.	Model behaviour plus an externally proposed mapping to a human construct.	Construct, subject	Explanation of which aspects of the human construct are preserved, absent or transformed.

Table 1: Analytical framework for interpreting bias-related findings in LLM evaluation.

measurement work in fairness and validity-threat analyses in empirical software engineering, which share a concern with how constructs are operationalized, how measurements support claims and how findings generalize.

The three mismatches discussed above identify points where an evaluation result can be interpreted more strongly than the evidence warrants. *Construct mismatch* cautions against treating model-side measures as direct instances of human constructs (Jacobs and Wallach, 2021). *Subject mismatch* cautions against inferring human-like mechanisms from behavioural similarity alone (Cuskley et al., 2024). *Evaluation-context mismatch* cautions against generalizing from prompt-conditioned outputs to broader claims about behaviour (Hida et al., 2025). Building on these cautions, we distinguish four interpretations of bias-related findings: *distributional association*, *observable behaviour*, *normative harm*, and *psychological analogy*.

Table 1 presents the analytical framework, linking each interpretation to typical model-side evidence, relevant mismatches, and the additional support needed to justify the claim. The framework is not intended to replace existing empirical evaluation methods with a new metric, benchmark or standardized evaluation protocol. Rather, it targets the inferential step connecting an empirical measurement to a bias claim. While existing practices often advise caution when interpreting LLM bias results, our framework specifies what kind of caution is needed, which type of claim is warranted by which type of evidence, and what additional support is required for stronger claims.

Distributional association refers to stronger model-side associations among groups, attributes or outcomes, measured through likelihood contrasts, embedding similarities or association scores (Gallegos et al., 2024; Bai et al., 2025; Lou and Sun, 2026). The result supports a limited claim that the model associates some terms, groups or outcomes more strongly than others under the chosen operationalization. This claim remains close to the measured quantity, but still requires justification that the measured association is a meaningful proxy for the intended construct. By itself, it does not show how the model will behave in interaction, whether the association will cause harm, or whether it reflects a human-like cognitive mechanism (Gawronski et al., 2022; Jacobs and Wallach, 2021).

Observable behaviour refers to bias expressed in model outputs under specific evaluation conditions, like stereotype-consistent generations, framing-sensitive answers, or decision changes after biased cues (Gallegos et al., 2024; Echterhoff et al., 2024). Observations are tied to the prompting setup, model version, decoding choices, and context (Hida et al., 2025; Holtzman et al., 2020). This supports a context-bound behavioural interpretation in which the model produces biased outputs in the tested setting. This interpretation requires attention to the evaluation context, because outputs are shaped by prompts, model versions, decoding choices, refusal behaviour and interaction context in which the behaviour was observed (Hida et al., 2025; Holtzman et al., 2020). Stronger behavioural claims therefore require evidence that the pattern is robust across relevant prompts, models, decoding settings or task variants.

Normative harm refers to model behaviour connected to individual or social consequences. Here, the interpretation moves beyond the existence of a pattern to whether it might reinforce inequality, degrade decisions or adversely affect users, groups or institutions in a particular context. For example, demographic disparity in a simulated hiring recommendation may matter because similar behaviour in deployed LLMs could reinforce unequal treatment (Blodgett et al., 2020). Harm claims require more than detecting a pattern. They require an account of the setting, affected users/groups, and consequences that follow from the model output (Blodgett et al., 2020; Gallegos et al., 2024).

Psychological analogy refers to treating model behaviour as meaningfully similar to a human cognitive or social-cognitive bias (Bai et al., 2025; Echterhoff et al., 2024; Lou and Sun, 2026). This is a stronger interpretation than an output pattern as it treats the model behaviour as understandable through a human-derived construct such as implicit bias, stereotype activation, anchoring, or confirmation bias. Such claims may be useful when carefully formulated, but they require explicit theoretical justification. In particular, researchers should specify which aspects of the human construct are preserved, absent, or transformed, and why the model-side observable supports the analogy (Gawronski et al., 2022).

The framework can be used as a reporting and interpretation checklist. It asks researchers to make explicit: (i) the human-derived construct being invoked; (ii) the model-side observable used to operationalize it; (iii) the evaluation context in which the observable is produced; and (iv) the strongest type of claim the design can support. As a result, the framework may leave the empirical measurement unchanged while changing the scope and strength of the conclusion drawn from it.

The framework does not make the classification of bias claims fully objective. Claims about harm and psychological analogy necessarily involve theoretical and contextual judgment. Its purpose is instead to make the basis for such judgments explicit and comparable across studies. Researchers should begin from the model-side observable and identify the strongest interpretation for which the required support is available. If the evidence consists of likelihood contrasts, embedding similarities or association scores, the warranted interpretation is closest to distributional association. If the evidence consists of generated outputs, rankings, recommendations, classifications or answer shifts under specified evaluation conditions, it may support observable behaviour, especially when robustness across prompts, models, decoding settings or task variants is shown. If the claim concerns harm, the evaluation must additionally specify a use context, affected groups and plausible consequences. If the claim invokes a human-derived construct, the evaluation must justify the mapping between the human construct and the model-side observable.

The IAT-inspired case provides a concrete example of how the framework changes the interpretation of an existing evaluation. Bai et al. (2025) show that LLMs can appear explicitly unbiased in direct responses while still exhibiting biased associations under association-based measures. Our framework does *not dispute this empirical finding, but it refines the conclusion that follows from it*. If the measured evidence is a likelihood or association difference between stereotype-consistent and -inconsistent group–attribute pairings, the most direct interpretation is *distributional association*. Interpreting the same results as biased model *behaviour* requires additional evidence that the associations shape generated outputs, ranking recommendations or other task behaviours. Interpreting it as a *normative harm* requires connecting the pattern to a specific use context, affected groups and plausible consequences. Interpreting it as human-like *implicit bias* requires a *psychological analogy* between the model-side observable and the human construct. The framework therefore changes the conclusion from a broad claim that the model has implicit bias to a more precise claim about what the evidence supports and what additional support would be needed for stronger interpretations.

These interpretations mark different inferential endpoints, including what the model seems to associate, what it produces under evaluation conditions, what consequences that behaviour may have in context, and whether it is meaningfully analogous to a human cognitive or social-cognitive bias[3].

5 Conclusions

Human-derived bias constructs offer valuable resources for evaluating LLMs beyond aggregate performance metrics and overt bias queries. However, these constructs do not transfer to LLM evaluation unchanged. When operationalized through token probabilities, generated outputs, rankings or answer shifts, the relation between evidence and interpretation becomes indirect. This paper argued that such transfer creates three sources of mismatch: construct, subject and evaluation-context. Mismatches do not invalidate psychological paradigms as tools for LLM evaluation, but show why claims about model bias should be stated more precisely. By distinguishing distributional association, observable behaviour, normative harm and psychological analogy, our framework shapes bias evaluation not only as a measurement task, but also as a problem of scientific interpretation. As LLMs become embedded in daily life tasks, NLP research needs not only better probes, but also clearer accounts of what those probes can and cannot reveal.

[3] Appendix C illustrates the same logic with an anchoring-style evaluation

Limitations

We acknowledge a number of limitations. First, it is a conceptual and literature-based contribution rather than an empirical evaluation. We do not test whether specific LLM bias probes produce stable results across models, prompts, languages or deployment settings. Instead, our goal is to clarify the kinds of interpretations that such probes can support. Future work could empirically examine how existing evaluations move between distributional, behavioural, normative and psychological claims, and whether those claims are warranted by the evidence provided. Second, the paper focuses on two broad families of human-derived constructs, social-cognitive bias constructs and cognitive biases in judgment and reasoning. This scope does not include all uses of psychological concepts in LLM evaluation, such as personality, emotion, moral judgment, theory of mind or pragmatic reasoning. The analytical framework may be useful for these areas as well, but additional work would be needed to adapt it to constructs with different theoretical histories, operationalizations and evidential standards. Finally, the four interpretations proposed here are not intended as an exhaustive taxonomy of model bias. Distributional association, observable behaviour, normative harm, and psychological analogy capture recurring ways in which bias-related findings are interpreted, but other distinctions may be needed for specific domains, modalities or applications. In particular, normative harm depends on social context, affected communities and deployment conditions, which cannot be fully specified by an abstract analytical framework alone.

Ethical Considerations

This paper does not introduce new datasets, model evaluations, or human-subject experiments. Its ethical concern is therefore primarily interpretative. One risk of imprecise claims about LLM bias is that they may anthropomorphize models, obscure the mechanisms through which harms arise, or overstate what particular evaluations demonstrate. Conversely, overly narrow interpretations may understate risks when biased outputs are likely to affect users, groups or society in deployment settings.

A further ethical consideration concerns the use of human-derived constructs such as implicit bias, stereotype activation or cognitive bias. These terms carry theoretical and social significance from psychology and social science. When applied to LLMs, they should be used in ways that avoid misleading equivalences between human cognition and model behaviour. Clearer distinctions between distributional association, observable behaviour, normative harm and psychological analogy can support more responsible communication of bias findings to researchers, developers, policymakers and affected communities.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF): 10.55776/COE12.

Declaration of Use of Generative AI Tools

Generative AI tools were used solely to assist with grammar correction and language polishing; all conceptual content, analysis, and final decisions remain the authors' responsibility.

References

Chittaranjan Andrade. 2018. Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian journal of psychological medicine*, 40(5):498–499.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5454–5476.

Lee Anna Clark and David Watson. 2019. Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*, 31(12):1412.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179.

Bertram Gawronski, Alison Ledgerwood, and Paul W Eastwick. 2022. Implicit bias≠ bias on implicit measures. *Psychological inquiry*, 33(3):139–155.

Tomer Geva, Ariel Goldstein, Eran Lary, and Coral Levy. 2025. Do llms exhibit human-like cognitive biases? a large-scale systematic evaluation. *A Large-Scale Systematic Evaluation (September 17, 2025)*.

Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2025. Social bias evaluation for large language models requires prompt variations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14507–14530, Suzhou, China. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

Jiaxu Lou and Yifan Sun. 2026. Anchoring bias in large language models: An experimental study. *Journal of Computational Social Science*, 9(1):11.

Simon Malberg, Roman Poletukhin, Carolin M Schuster, and Georg Groh. 2025. A comprehensive evaluation of cognitive biases in llms. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 578–613.

Mario Mina, Valle Ruiz-Fernández, Júlia Falcão, Luis Vasquez-Reina, and Aitor Gonzalez-Agirre. 2025. Cognitive biases, task complexity, and result interpretability in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1767–1784, Abu Dhabi, UAE. Association for Computational Linguistics.

Jeremy K Nguyen. 2024. Human bias in AI models? anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance*, 43:100971.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. 2024. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024(1):7115633.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models display human-like social desirability biases in big five personality surveys. *PNAS nexus*, 3(12):pgae533.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629, Miami, Florida, USA. Association for Computational Linguistics.

Shweta Soundararajan and Sarah Jane Delany. 2024. Investigating gender bias in large language models through text generation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 410–424, Trento. Association for Computational Linguistics.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2025. Cognitive biases in large language models: A survey and mitigation experiments. In *Proceedings of the 40th ACM/sigapp symposium on applied computing*, pages 1009–1011.

Yoshiki Takenami, Yin Jou Huang, Yugo Murawaki, and Chenhui Chu. 2025. How does cognitive bias affect large language models? a case study on the anchoring effect in price negotiation simulations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4481–4498, Suzhou, China. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Irene CE van Blerck, Edirlei Soares de Lima, Margot ME Neggers, and Toon Calders. 2025. Unveiling gender bias in llm-generated hero and heroine narratives. *Entertainment Computing*, 55:100972.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 82232–82251. PMLR.

Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering*, volume 236. Springer.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Ruifang He, and Yuexian Hou. 2025. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1–12, Vienna, Austria. Association for Computational Linguistics.

A Examples of the Inferential Gap in Existing Bias Evaluations

This Appendix provides examples of how the inferential gap discussed in the paper appears in existing LLM bias-evaluation studies. The goal is *not to dismiss these studies or dispute their empirical findings*. Rather, the examples show *how human-derived constructs can support useful measurements while also inviting stronger interpretations that require additional support*.

Political bias and self-perception. Rutinowski et al. (2024) evaluate ChatGPT using political-compass questionnaires, political-affiliation questions and personality-related instruments. These instruments provide evidence about how the model responds under particular questionnaire conditions. However, stronger interpretations, such as claims that the model holds political views, perceives itself as having certain personality traits, or even that it has a stable personality type, depend on whether human-oriented political and personality instruments can validly measure such constructs in LLMs. This illustrates construct mismatch and subject mismatch, as political identity, self-perception and personality are human constructs involving social, experiential and psychological dimensions.

Social desirability bias. Salecha et al. (2024) study social desirability bias in LLM responses to Big Five personality surveys. Their findings show that model responses shift toward socially desirable trait profiles when the evaluation context becomes identifiable. This is evidence about response behaviour under a psychometric evaluation setting. At the same time, interpreting the pattern as human-like social desirability bias requires specifying which parts of the human construct are preserved or transformed in the LLM setting. In humans, social desirability is often linked to self-representation, impression management and sensitivity to social norms. In LLMs, similar response patterns may instead reflect training data, instruction following, alignment procedures or learned conventions.

Anchoring bias. Takenami et al. (2025) examine anchoring effects in LLM-based price negotiation simulations. Results show that model outputs and negotiation outcomes shift under anchoring-style interventions. This provides evidence of context-sensitive behaviour in a simulated negotiation setting. A stronger interpretation that LLMs are in-

fluenced by anchoring in the same sense as humans requires an explicit mapping between the human cognitive-bias construct and the model-side behaviour.

These examples show why the inferential gap is not merely hypothetical. Existing evaluations can produce informative findings while still requiring care in how those findings are translated into claims about model beliefs, personality, social desirability, cognitive bias, harm or human-like psychological processes.

B Example Distinguishing Two Families of Bias Probes

This Appendix illustrates the distinction between the two families of bias probes discussed in Section 2. Both families transfer human-derived constructs to model-side observables, but they differ in the construct invoked, the evidence collected and the interpretation that is most directly warranted.

Social-cognitive bias probe. An IAT-inspired probe tests whether an LLM produces stronger associations between social groups and stereotyped attributes, for example whether gendered terms are more strongly associated with career- or family-related terms. The human-derived construct comes from social cognition and prejudice research, where the IAT uses response-time differences to infer relative associations between concepts (Greenwald and Banaji, 1995; Greenwald et al., 1998). In an LLM adaptation, the model-side observable may instead be a likelihood contrast, embedding similarity, ranking preference or generated continuation. Therefore, the most direct interpretation is a claim about *distributional association*, unless further evidence shows that the association affects generated outputs, produces harms in a specific use context or is analogous to human implicit bias.

Cognitive-bias probe. An anchoring-style probe asks whether an LLM's answer shifts toward an irrelevant numerical cue. The human-derived construct comes from work on judgment and reasoning, where anchoring refers to the influence of an initial value on subsequent estimates or decisions (Tversky and Kahneman, 1974). In an LLM adaptation, the model-side observable may be an answer shift across controlled prompt variants, such as different estimates after a low or high anchor. The most direct interpretation is a claim about *observable behaviour* under the tested prompt conditions. A stronger claim about human-like anchoring requires showing that the anchor is task-irrelevant, that the task has a meaningful target or baseline, and that the model systematically shifts toward the anchor across controlled variants.

This contrast shows why the two probes should not be collapsed into a single notion of *LLM bias*. In the first case, the central evidential issue is whether a model-side association is a meaningful proxy for a social-cognitive construct. Whereas in the second case, the central issue is whether an output shift reflects a cognitive-bias pattern rather than general prompt sensitivity. In both cases, stronger claims about harm or psychological analogy require additional support beyond the initial model-side measurement.

C Applying the Framework to an Anchoring-Style Evaluation

Consider an evaluation in which a model gives a higher estimate after exposure to an arbitrary numerical anchor. The direct finding is that the model output shifts under a controlled prompt variation. Under our framework, this supports an *observable behaviour* claim, namely that the model is sensitive to an anchoring-style cue under the tested conditions.

A stronger claim about human-like anchoring requires additional support. Researchers should specify what is meant by anchoring in the LLM setting, why the anchor is task-irrelevant, whether the task has a meaningful target answer or baseline, and whether the model systematically shifts toward the anchor across controlled variants. Without this mapping, the result is better described as context-sensitive model behaviour than as human-like anchoring.

A harm-based interpretation would require further support of a different kind. The output shift would need to be connected to a concrete use setting, affected users or groups, and plausible consequences, such as distorted salary estimates in compensation support, biased evaluations in financial decision support or unfair outcomes in online marketplaces. Thus, the framework changes the conclusion from a broad claim that the model has anchoring bias to a more precise claim about context-sensitive model behaviour, unless additional evidence supports the psychological or harm-based interpretations.