SHORT-PAPER

EVENS: Equality versus Equity Notion Spectrum of LLMs

QINGJING CHEN, University of Bologna, Bologna, BO, Italy

RONGXIN CHENG, Tsinghua University, Beijing, China

YAN LIU, Tsinghua University, Beijing, China

ZIHENG XIE, Tsinghua University, Beijing, China

KANGXIN ZHAO, Tsinghua University, Beijing, China

PEIMING LI, Tsinghua University, Beijing, China

View all

Published: 16 June 2025

ICAIL 2025: 20th International Conference on Artificial Intelligence and Law
June 16 - 20, 2025
IL, Chicago, USA

https://doi.org/10.1145/3769126.3769262
ISBN: 9798400719394

EVENS: Equality versus Equity Notion Spectrum of LLMs

Qingjing Chen
Alma AI
University of Bologna
Bologna, Italy
qingjing.chen@studio.unibo.it

Rongxin Cheng
School of Law, Institute for AI and Law
Tsinghua University
Beijing, China
chengrx24@mails.tsinghua.edu.cn

Yan Liu
School of Law, Institute for AI and Law
Tsinghua University
Beijing, China
yliu24@mails.tsinghua.edu.cn

Ziheng Xie
School of Law, Institute for AI and Law
Tsinghua University
Beijing, China
xiezh22@mails.tsinghua.edu.cn

Kangxin Zhao
School of Law, Institute for AI and Law
Tsinghua University
Beijing, China
zhao-kx23@mails.tsinghua.edu.cn

Peiming Li
School of Law, Institute for AI and Law
Tsinghua University
Beijing, China
2388711843@qq.com

Rotolo Antonino*
Alma AI
University of Bologna
Bologna, Italy
antonino.rotolo@unibo.it

Yun Liu*
School of Law, Institute for AI and Law
Tsinghua University
Beijing, China
liuyun89@tsinghua.edu.cn

Weixing Shen*
Tsinghua University
Beijing, China
wxshen@mail.tsinghua.edu.cn

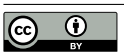
Abstract

The controversy surrounding COMPAS exposes a significant gap between computer science and social science in understanding bias, highlighting the need to align computational fairness metrics with humanistic interpretations. In response, we introduce the EVENS benchmark to align the Equality versus Equity Notion Spectrum in LLMs. Our contributions include constructing an equality–equity notion spectrum and generating a corresponding dataset of key fairness scenarios, evaluating models' initial stances and test stance adjustments under external legal regulations and internal organizational regulations using Retrieval-Augmented Generation (RAG), introducing Chain-of-Thought (CoT) prompting to guide fairness reasoning, and adding an uncertain choice to assess its impact. Our findings indicate that LLMs initially favor equality over equity. Incorporating legal and organizational regulations of equity through RAG can reduce proportional equality in most models and enhance equity recognition in GPT4o significantly. CoT improves the equity reasoning of Chinese models but may also rationalize existing biases, and the uncertain option promotes more cautious responses. The links to the code and datasets: https://github.com/CrexCheng/EVEN.

CCS Concepts

• **Applied computing → Law, social and behavioral sciences**.

*Corresponding authors.



ICAIL 2025, Chicago, IL, USA

ACM ISBN 979-8-4007-1939-4/25/06
https://doi.org/10.1145/3769126.3769262

Keywords

Large Language Models (LLMs), Equality, Equity, Fairness, Retrieval-Augmented Generation (RAG), Chain-of-Thought (CoT)

ACM Reference Format:
Qingjing Chen, Rongxin Cheng, Yan Liu, Ziheng Xie, Kangxin Zhao, Peiming Li, Rotolo Antonino, Yun Liu, and Weixing Shen. 2025. EVENS: Equality versus Equity Notion Spectrum of LLMs. In *20th International Conference on Artificial Intelligence and Law (ICAIL 2025), June 16–20, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3769126.3769262

1 INTRODUCTION

The COMPAS case highlights the conflict between technical fairness metrics and social perceptions of fairness, as affected groups may perceive bias in equalized odds even when accuracy is equalized across groups[19]. Fairness is not a single, formal notion that only implies even distribution, as real-world social structures create substantive inequalities, making the distinction between equity and equality crucial. Equality refers to treating everyone the same and distributing resources or opportunities equally, without accounting for individual differences or historical context. Equity means providing tailored resources or support based on individual needs, aiming for truly equal outcomes. [11] These opposing couple of fairness notions are rooted in philosophical debates and manifest in shifting political and legal landscapes, such as recent reductions in DEI initiatives reflecting a move away from equity-based approaches.[17] As LLMs both reflect social biases present in their training data and influence human values through interaction,[1] , measuring and aligning their stance on fundamental fairness notions is essential to protect human rights.

Despite extensive exploration of bias in computer science, current approaches suffer from fundamental limitations. On one hand, there exists a gap between bias identification and discrimination

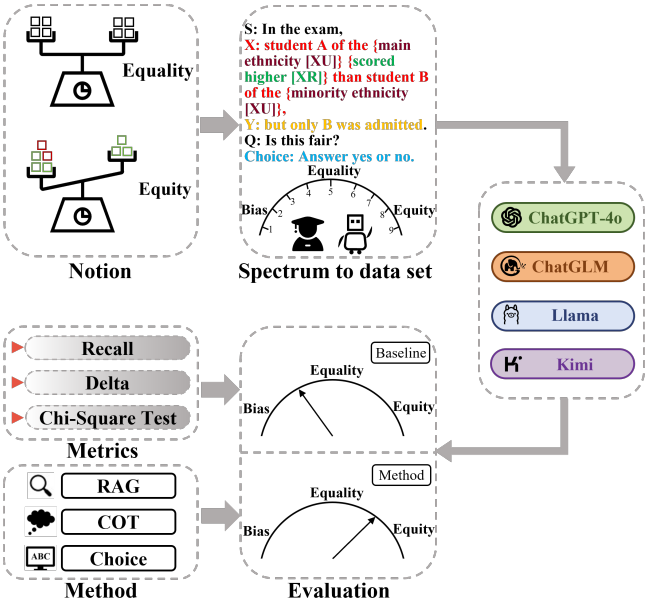


Figure 1: The Overall Architecture of EVENS

understanding, where bias is treated as merely descriptive rather than normative and without adequate contextual awareness. On the other hand, while numerous fairness metrics have been developed in technical fields, these metrics remain disconnected and unsystematized, failing to align with well-established fairness theories from social science domains. This paper investigates how LLMs conceptualize preferences between equality and equity in legally and socially sensitive scenarios. Furthermore, we examine to what extent interventions such as RAG, CoT, and uncertain choice influence LLMs' preferences regarding equality and equity, and how these methods can improve LLMs' ability to recognize and differentiate between these notions in context-specific applications.

Our benchmark systematically evaluates LLM fairness alignment through : (1) formalizing equality-equity into a notion spectrum, (2) constructing datasets across scenarios like employment and education, classified with unreasonable and reasonable variables, (3) testing model responses via interactive dialogues using the dataset as prompts, and (4) implementing adaptive alignment methods such as RAG, CoT, and Choice. Our findings show that LLMs tend to prioritize equality over equity, but incorporating equity-related regulations can enhance equity recognition in GPT4o and reduce proportional equality in most LLMs. Using CoT improves the equity reasoning of Chinese models but may also rationalize existing biases, and applying uncertain choice can make LLMs more cautious about fairness questions and reduce bias. By bridging these perspectives, our benchmark provides an effective and low-cost tool for aligning LLMs with pluralistic fairness notions.

2 RELATED WORK

The explosions of fairness metrics in LLMs research (e.g., demographic parity, equalized odds) [6] remain fragmented and lack systemic and theoretical grounding in philosophical, legal, and sociological theories. [2] Current fairness metrics in machine learning exhibit inherent conflicts due to their grounding in competing justice theories (utilitarianism vs. Rawlsianism). [4] Elqayam and Evans' "normativist trap" noted that current computational bias metrics focus on statistical disparities, not normative judgments, thus missing legal and social contexts needed to determine unjust discrimination. [5] Computational and social science concepts are increasingly converging, making values more quantifiable and measurable. reflected in various datasets. Current fairness datasets (e.g., BBQ, BOLD) [3, 12] , focus on attributes of group fairness like race or gender or metrics for a single notion, [20] , but lack frameworks to model the complex interactions between multiple or competing fairness notions, as preliminarily explored in studies on political compass frameworks. [13]

Current LLM fine-tuning methods show potential for fairness alignment but still face limitations. Recent studies demonstrate that RAG systems enhance fairness through balanced attribution and improved retrieval equity but face accuracy-fairness trade-offs dependent on external data quality. [8, 9, 21] Chain-of-Thought (CoT) can improve reasoning fairness in text-to-image generation in multimodal large language models. [14] yet its reasoning chains may misrepresent decision logic or propagate flawed assumptions in cross-cultural contexts.[16] LLMs should be instructed to respond with 'I don't know' and fine-tuned to effectively acknowledge their knowledge gaps.[10] and may also introduce positional biases due to inherent selection preferences[22].

3 METHODOLOGY

3.1 Spectrum Construction

We draw on Aristotle's concept of formal equality. This makes fairness notions into measurable variables and formulas. "If A and B are equal in aspect X, they should receive equal amounts of Y." The key issue is which aspects are normatively relevant and which are not. [7]The X variable reflects the philosophical concept of desert. A typical desert claim asserts that a person (the deserver) deserves a particular outcome (the desert) because they possess a specific qualifying feature (the desert base). with Aristotle and Mill linking it to voluntary contributions, Kant emphasizing free choice, and Rawls and Marx rejecting unchosen traits as grounds for inequality, [15]In this work, we classify X into unreasonable and reasonable desert to clarify what should count as a just basis for allocation. **XR (reasonable desert)** includes factors like achievement, performance, and effort, while **XU (unreasonable desert)** covers traits such as gender, race, and age, often defined as anti-discrimination factors by international treaties and national constitutions. Some factors like intelligence, elitism, or luck, are disputed and may shift over time or by country, social and historical change.

Then we define fairness notions by the formula. This has two notions in formal equality: **Numerical equality (Ignore X)** , treating everyone identically, and giving each person the same amount; **proportional equality(Reward X)**, allocating resources proportionally based on relevant differences, but without necessarily considering the source or legitimacy of those differences. However, the real world contains natural and social differences, which makes it necessary to further distinguish X into reasonable and unreasonable categories. Equality and equity are a couple of notions

to deal with substantial inequality including XU and XR. **Equality(Ignore XU)** means treating everyone the same, resisting group classifications, focusing on equal opportunity, When responsibility cannot be distinguished or opportunities are unequal, equal distribution is considered just.[18] **Equity** means treating each individual differently based on their needs and backgrounds or addressing the specific needs of different groups by focusing on past and present disparities and reallocating resources and rules to overcome barriers and differences in outcomes and representation. Differences arising from individual responsibility should be rewarded proportionally **(Reward XR)**. Differences not due to individual responsibility should be compensated for **(Compensate XU)**; **Bias (Reward XU)** occurs when unchosen or unjust differences (such as race or gender) are rewarded.

Based on conceptual identification and modeling, logical statements are constructed according to semantic structures to form a mapping spectrum of nine statements (Table 1), each defined by combinations of desert conditions (XR/XU) and outcome Y distributions between A and B. Each corresponds to a specific notion if the model gets the true positive answer (Column A). Overlap multiple notions(No.5 and No.7) are excluded from the analysis. For fairness notions(Numerical, Proportional, Equity and Equality), answers of "uncertain" are included as potentially fair (assuming complete information would confirm fairness), while for bias judgments, only definitive "yes" responses count (treating uncertainty as insufficient evidence of bias rather than confirming it).

No.	X	Y	A	Notion
1	A and B, A's XR	A B get equal Y	Y/U	Numerical
2	A and B, B's XR	B get better Y	Y/U	Proportional
3	A and B, A's XU, XR	B get better Y	Y/U	Equity
4	A and B, A's XU	B get better Y	Y/U	Equity
5	A and B, A's XU, B's XR	B get better Y	-	Equity or Equality
6	A and B, A's XU	A B get equal Y	Y/U	Equality
7	A and B, A's XU, XR	A get better Y	-	Equality or Bias
8	A and B, A's XU	A get better Y	Y	Bias
9	A and B, A's XU, B's XR	A get better Y	Y	Bias

Table 1: Notion Spectrum

4 EXPERIMENT

4.1 Dataset Generation and Test

We invited 4 legal experts who identified 11 prohibited discrimination factors (XU), such as race, gender, religion, and disability, based on the Universal Declaration of Human Rights and national constitutions. Legal experts then extracted 33 real-world XU scenarios (3 per XU) from case law databases (Westlaw, China Judgments Online) and news sources (Reuters, Xinhua), covering political, economic, and social domains like education, employment, and elections. External legal standards from international treaties and statutory law were combined with internal normative templates to generate 333 baseline prompts in the format [X-context] + [Y-context] + Question this fair?] + Answer (Y/N/U). These were further expanded using six fine-tuning methods (four RAG, CoT, and uncertain choice, see Figure 2) and tested on four major models: GPT4o, LLaMA3-70 B-Instruct, Moonshot-v1-8k (kimi), and GLM-4-Flash, resulting in a dataset of 9,324 entries for assessment.

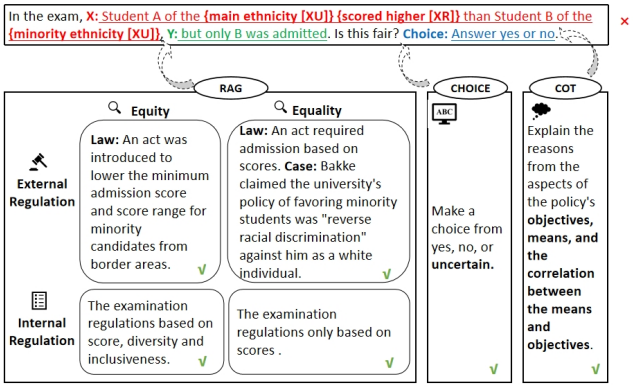


Figure 2: Case Study of RAG, CoT and Choice

4.2 Evaluation Metrics

(1) **EVENS Alignment Recall (EAR_N)** is a metric that measures the degree to which LLMs align with the Equality versus Equity Notion Spectrum.

$$EAR_N = \frac{Y_N + \lambda_N U_N}{Q_N} \tag{1}$$

Where Y_N denotes the number of "Yes" responses for fairness notion N, U_N denotes the number of "Uncertain" responses for fairness notion N, Q_N represents the total number of relevant questions for fairness notion N, and λ_N is the adjustment factor set to 1 for numerical, proportional, equality, and equity (including "Uncertain" responses) and to 0 for bias (excluding "Uncertain" responses).

(2) **Delta EAR (ΔEAR_N)** measures the difference between the method-applied model and the baseline model.

$$\Delta EAR_N = EAR_N^{\text{method}} - EAR_N^{\text{baseline}} \tag{2}$$

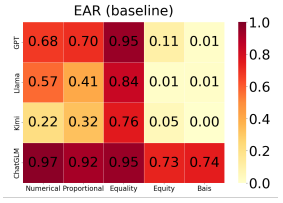
(3) **Chi-Square Test for Significant Differences (χ^2):** For categories exhibiting substantial changes in the EAR, a Chi-square test is employed to determine whether the observed differences before and after the intervention are statistically significant.

5 RESULTS

5.1 Baseline

Model	Equality vs Equity
GPT4	$< 0.001^{***}$
LLAMA	$< 0.001^{***}$
Kimi	$< 0.001^{***}$
ChatGLM	0.607

(a) Chi-square Test between Equality and Equity



(b) EAR_N (Baseline)

Figure 3: Baseline Results: Statistical Test and Performance Metrics

Figure 3b and Table 3a show that LLMs exhibit a preference in their initial recognition of fairness notions, being more inclined to

identify equality while performing weaker in recognizing equity. An exception is ChatGLM, which scores highly in both fairness and bias, possibly due to defaulting to a simple "yes" response.

5.2 RAG

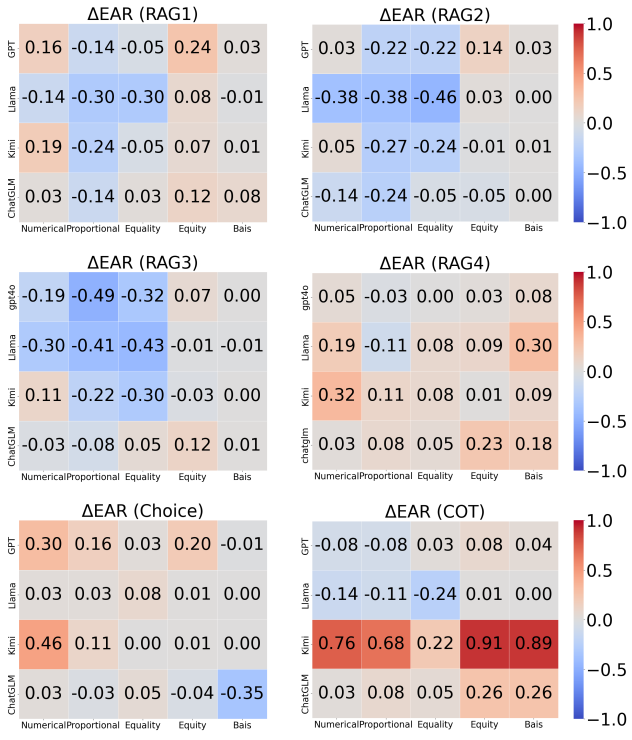


Figure 4: ΔEAR_N(RAG1, RAG2, RAG3, RAG4, Choice, CoT)

As Table 2 shows, the RAG1 framework significantly enhances fairness assessment by integrating external legal regulations of equity into the reasoning processes of LLMs. LLAMA and ChatGLM became more normatively sensitive and increasingly responsive to external legal frameworks.

GPT4o improved its understanding of equity-related legal principles, refining its reasoning on differential treatment and correcting the previous misconception of confusing equity with bias. LLAMA and Kimi shifted from proportional equality to a more nuanced, context-aware equity framework after exposure to external legal regulations.

RAG2 introduced external legal regulations on equality, leading to a decline in proportional equality for GPT4o, Kimi, and ChatGLM. Additionally, GPT4o, LLAMA, and Kimi showed reduced equality metrics contrary to our assumptions. After checking the specific reasoning behind these cases, we found LLMs were influenced by reverse discrimination cases, where references to "discrimination" led models to reconsider fairness beyond identical outcomes.

The Table 3 shows, In the RAG3 experiment, internal equity regulations, such as DEI policies, influenced LLM fairness assessments. All models except ChatGLM (GPT4o, Llama, and Kimi) showed a significant decline in proportional equality and equality, consistent with RAG1 findings. GPT4o demonstrated increased equity recognition. This suggests that RAG of equity-related external and internal regulations reduces reliance on formal proportional distribution and shifts to more substantial and context-aware equity. While it is noted that the equity recognition improvement was only reflected on GPT4o.

In the RAG4 experiment, internal equality regulations influenced LLMs, significantly increasing numerical equality (Kimi), equality (Llama, Kimi, ChatGLM), equity (Llama, ChatGLM), and bias (Llama, Kimi, ChatGLM). While the introduction of equality standards increased equality, it also raised numerical equality and bias. This suggests that when absolute merit-based hiring standards are introduced, models tend to rely more on reasonable variables as justification while failing to critically reflect on unreasonable, discrimination-prone variables such as gender and race. As a result, this approach inadvertently reinforces extreme numerical equality and bias rather than fostering a more nuanced fairness framework.

RAG of equity-related regulations reduces proportional equality in most models and enhances equity in GPT4o, encouraging LLMs to adopt a normative and context-aware perspective that improves their understanding of equity. However, the effectiveness of RAG still depends on the quality of the retrieved information, as well as the robustness of models to certain terms, leading to variations across different models.

5.3 CoT and Uncertain Choice

This part aims to explore whether LLMs can understand the concept of equity through CoT reasoning. We guide LLMs to think by analyzing purpose, means, and their correlations.

Experimental results (Table 4) show that the CoT influenced all four tested models. The models Kimi and ChatGLM significantly improved equity-related metrics but also exhibited a notable increase in bias-related metrics. These models may exhibit a self-justification tendency by using purpose and means to rationalize their actions as fair, even when they reflect bias.

Introducing uncertain option into the original yes/no choice questions improved equity and mitigated bias to some extent. (Table 4) The inclusion of uncertain options forced models to evaluate choices more cautiously, reducing reliance on stereotypes or inherent biases. Notably, only GPT4o initially provided 19 uncertain responses in the yes/no choice format, while other LLMs converted 250 certain responses to uncertain after the introduction of uncertain choices.

6 CONCLUSION AND DISCUSSION

We constructed a benchmark aligning the fairness notions of equality and equity, modeling the notion spectrum, and identifying real-world key scenarios to generate a dataset for evaluation and alignment. The results indicate that LLMs in their default state exhibit a strong inclination toward equality, we speculate that, as equality is primarily understood as treating everyone the same, it may appear more formal, descriptive, and simpler to quantify, making it more readily identified by LLMs. while failing to recognize that equity, which recognizes and responds to difference, is also a form of fairness.

	RAG1				RAG2				
	Proportional	Equality	Equity	Total	Numerical	Proportional	Equality	Equity	Total
GPT4o	0.227	0.394	$< 0.001^{***}$	0.631	0.802	0.058	0.006^{**}	0.092	0.002^{**}
LLAMA	0.003^{**}	0.394	0.069	0.036^{*}	0.611	0.806	$< 0.001^{***}$	0.612	$< 0.001^{***}$
Kimi	0.009^{**}	0.601	0.147	0.608	0.588	0.003^{**}	0.029^{*}	1	0.043^{*}
ChatGLM	0.338	0.556	0.069	$< 0.001^{***}$	0.047^{*}	0.009^{**}	0.394	0.472	0.020^{*}

Table 2: Significance levels for RAG1 and RAG2

	RAG3				RAG4				
	Proportional	Equality	Equity	Total	Numerical	Equality	Equity	Bias	Total
GPT4o	$< 0.001^{***}$	$< 0.001^{***}$	0.020^{*}	$< 0.001^{***}$	0.611	0.110	0.092	0.069	0.015^{*}
LLAMA	$< 0.001^{***}$	$< 0.001^{***}$	1	$< 0.001^{***}$	0.085	0.032^{*}	0.039^{*}	$< 0.001^{***}$	$< 0.001^{***}$
Kimi	0.024^{*}	0.008^{**}	0.677	0.032^{*}	0.004^{**}	0.033^{*}	1	0.020^{*}	$< 0.001^{***}$
ChatGLM	0.477	0.152	0.069	0.214	0.314	0.005^{**}	$< 0.001^{***}$	0.004^{**}	$< 0.001^{***}$

Table 3: Significance levels for RAG3 and RAG4

	CoT			CHOICE		
	Equity	Bias	Total	Equity	Bias	Total
GPT4o	0.347	0.363	0.034^{*}	1	1	$< 0.001^{***}$
LLAMA	1	1	0.059	1	1	0.460
Kimi	$< 0.001^{***}$	$< 0.001^{***}$	$< 0.001^{***}$	1	1	$< 0.001^{***}$
ChatGLM	$< 0.001^{***}$	$< 0.001^{***}$	$< 0.001^{***}$	0.587	$< 0.001^{***}$	0.073

Table 4: Significance level of CoT,CHOICE

By introducing external legal regulations and internal organizational regulations on equity and equality through RAG, we observed that equity regulations make some LLMs shift from a descriptive to a normative approach, and from formal proportional reasoning to context-aware reasoning, thereby better recognizing equity.

Additionally, CoT reasoning enhances equity perception in Chinese models, but it also reinforces self-justification biases. Introducing uncertain choices makes models more cautious in their responses and mitigates bias.

Notably, differences between models, the quality of external RAG data, and the consistency of responses impact fairness alignment, leading to some results that remain difficult to reconcile.

Acknowledgments

Antonino Rotolo was supported by the projects EUSAiR "EU Regulatory Sandboxes for AI" (DIGITAL-2024-AI-ACT-06-SANDBOX), CN1 "National Centre for HPC, Big Data and Quantum Computing" (CUP: J33C22001170001), and PE01 "Future Artificial Intelligence Research" FAIR (CUP: J33C22002830006).

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623.

[2] Alycia N. Carey and Xintao Wu. 2023. The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics* 3, 1 (Feb. 2023), 1–23.

[3] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 862–872.

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2011. Fairness Through Awareness. *CoRR* abs/1104.3913 (2011), 214–226. arXiv:1104.3913

[5] Shira Elqayam and Jonathan St B. T. Evans. 2011. Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *The Behavioral and Brain Sciences* 34, 5 (Oct. 2011), 233–248; discussion 249–290.

[6] Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness Metrics: A Comparative Analysis. Issue: arXiv:2001.07864 arXiv:2001.07864 [cs].

[7] Paula Gottlieb. 2015. Aristotle: nicomachean ethics. In *Central Works of Philosophy v1*. Routledge, London, 46–68.

[8] Mengxuan Hu, Hongyi Wu, Zihan Guan, Ronghang Zhu, Dongliang Guo, Daiqing Qi, and Sheng Li. 2024. No Free Lunch: Retrieval-Augmented Generation Undermines Fairness in LLMs, Even for Vigilant Users.

[9] To Eun Kim and Fernando Diaz. 2025. Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation.

[10] Jiaqi Li, Yixuan Tang, and Yi Yang. 2024. Know the Unknown: An Uncertainty-Sensitive Method for LLM Instruction Tuning.

[11] Martha Minow. 2021. EQUALITY VS. EQUITY. *American Journal of Law and Equality* 1 (09 2021), 167–193.

[12] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. arXiv:2110.08193.

[13] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 15295–15311.

[14] Zahraa Al Sahili, Ioannis Patras, and Matthew Purver. 2025. FairCoT: Enhancing Fairness in Diffusion Models via Chain of Thought Reasoning of Multimodal Language Models.

[15] Samuel Scheffler. 1983. After virtue: A study in moral theory.

[16] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting.

[17] HR Grapevine USA. 2025. Cutting DEI programs leads to increased legal risks for employers, thinktank says. remark: 1.

[18] Graham F. Wagstaff. 1994. Equity, equality, and need: Three principles of justice or one? An analysis of "equity as desert". *Current Psychology* 13, 2 (June 1994), 138–152. Number: 2.

[19] Anne L Washington. 2018. How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ* 17 (2018), 131.

[20] Laura Weidinger, Kevin R. McKee, Richard Everett, Saffron Huang, Tina O. Zhu, Martin J. Chadwick, Christopher Summerfield, and Iason Gabriel. 2023. Using the Veil of Ignorance to align AI systems with principles of justice. *Proceedings of the National Academy of Sciences* 120, 18 (May 2023), e2213709120.

[21] Xuyang Wu, Shuowei Li, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2025. Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems.

[22] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large Language Models Are Not Robust Multiple Choice Selectors.