INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 6 AND ISSUE 5 OF 2026

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

ARE LLM MODELS BIASED REGARDING CASTE STEREOTYPES IN THE INDIAN CONTEXT? - AN EMPIRICAL REVIEW AND TECHNO-LEGAL ANALYSIS OF AI BIAS MITIGATION FRAMEWORKS

AUTHOR - VISTAAR SINGH, STUDENT AT ATAL BIHARI VAJPAYEE SCHOOL OF LEGAL STUDIES, CSJM UNIVERSITY, KANPUR

BEST CITATION – VISTAAR SINGH, ARE LLM MODELS BIASED REGARDING CASTE STEREOTYPES IN THE INDIAN CONTEXT? - AN EMPIRICAL REVIEW AND TECHNO-LEGAL ANALYSIS OF AI BIAS MITIGATION FRAMEWORKS, *INDIAN JOURNAL OF LEGAL REVIEW (IJLR),* 6 (5) OF 2026, PG. 01-05, APIS – 3920 – 0001 & ISSN - 2583-2344.

ABSTRACT

Though banned by law, old rankings based on birth still shape who gets what in daily life across India. Trained on uneven data, artificial systems quietly mirror these inherited divides. Instead of questioning fairness, many tools accept biased inputs as normal. One inquiry probes whether machines treat people differently due to caste while using local tongues. Evidence gathered from peer-reviewed work and policy texts, current through early 2026, shows repeated links between low-status names and negative traits. High-caste labels tend to cluster around words like skillful or authoritative. These associations do not appear randomly; they echo historical power imbalances baked into digital forms. Regulatory efforts exist, yet their real-world impact remains limited so far. What appears neutral often carries forward long-standing exclusions. As the discussion winds down, attention turns to the necessity of binding regulations, external monitoring, context-specific protections, along with joint initiatives, so artificial intelligence does not deepen historical inequalities tied to caste across India.

1. Introduction

Spreading fast, large language models like ChatGPT from OpenAI, Google DeepMind's Gemini, Anthropic's Claude, along with homegrown versions such as Ola's Krutrim, have changed how many Indians engage with digital information. Starting with college admissions help, moving into hiring support, medical analysis, or legal guidance, they now shape key choices tied to fairness and life chances, choices backed by research (Bender et al., 2021[1003]). Behind this shift lies a striking fact: India ranks second worldwide in OpenAI usage, revealing both widespread adoption and growing need to examine real-world effects.

Caste appears embedded within language itself. Terms tied to kinship, place of origin, work roles; also how people speak to one another, often signal caste distinctions across Indian tongues (Jodhka & Newman, 2010[1004]). Because such markers fill everyday speech and writing, artificial systems learning from text naturally pick them up. Trained on vast corpora where these cues repeat, large language models begin mirroring social ranking. Over time, they start connecting particular last names with intellect, tie Dalit-sounding labels to menial

[1003] 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 (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922, accessed 24 Mar. 2026.

[1004] Jodhka, S.S., & Newman, K. (2010). In the Name of Globalisation: Meritocracy, Productivity, and the Hidden Language of Caste. Economic and Political Weekly, 45(7). https://www.jstor.org/stable/40276546, accessed 29 Mar. 2026.

jobs, or generate stories soaked in old notions of purity, all quietly reinforcing bias through code. What once lived in tradition now surfaces in algorithms.

Even though the issue demands attention, research on caste bias within AI fairness often overlooks it, centering instead on race, gender, and nationality, mainly in Western settings (Sambasivan et al., 2021[1005]; Birhane, 2021[1006]). Without such inquiry, critical gaps persist, given that caste-based inequity extends beyond society into economics and governance, shaping life chances over decades. When large language models absorb and echo caste-related assumptions, they may amplify entrenched disadvantages, quietly embedding exclusion more widely than before[1007].

This study tackles two connected inquiries. One question looks at whether large language models show detectable prejudice tied to caste within Indian language and societal settings. Another examines how India's legal and technological regulations currently handle detection, oversight, and reduction of such algorithmic slant, pinpointing key shortcomings. Drawing together scholarly articles and official guidelines up to April 2026, a fresh analysis emerges fit for academic review.

2. Sources of Data

Data for this study are drawn from two main categories:

- Empirical Data Sources (Secondary)

This study synthesizes findings from established and peer-reviewed empirical research focused on detecting caste-based bias in LLMs. Key sources include:

- DeCaste Dataset and Evaluation Framework[1008]: This system, presented at IJCAI 2025, provides quantitative and qualitative data on caste-related prejudices in LLMs. It includes results from:
  - Stereotype Word Association Test (SWAT): Analyzing LLM responses to names associated with different castes to identify stereotypical links.
  - Persona-based Scenario Association Test (PSAT): Examining LLM behavior in everyday situations involving identity-shaped characters across cultural, financial, learning, and governance dimensions.

  The data from DeCaste includes bias scores (ranging from -1 to +1) across nine leading LLMs, indicating the degree of stereotype alignment.
- Indian-BhED Dataset[1009]: This dataset comprises 229 pairs of English sentences designed to probe assumptions related to caste (Brahmin vs. Dalit) and religious identity. Data from this source includes the percentage of caste-based cases where LLMs favored expected cultural clichés (61% to 79%).
- Educational Support Disparity Analysis: Research by Gupta et al. (2026)[1010] provides qualitative empirical data through case studies demonstrating how LLMs adjust the complexity and depth of educational explanations based on perceived caste identity. This includes examples of differential information provision for Brahmin versus Dalit students.

[1005] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 315–328. https://doi.org/10.1145/3442188.3445896, accessed 20 Mar. 2026.

[1006] Abeba Birhane. (2021). Algorithmic injustice: a relational ethics approach. Patterns, 2(2), 100205. https://www.sciencedirect.com/science/article/pii/S2666389921000155, accessed 24 Mar. 2026.

[1007] Navigli, R., Conia, S., & Ross, B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM Journal of Data and Information Quality, 15, 2, Article 10.* https://doi.org/10.1145/3597307, accessed 20 Mar. 2026.

[1008] Vijayaraghavan, P., Vosoughi, S., Chiazor, L., Horesh, R., Abreu De Paula, R., Degan, E., & Mukherjee, V. 2025. DECASTE: unveiling caste stereotypes in large language models through multi-dimensional bias analysis. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI '25). Article 1100, 9899–9907. https://doi.org/10.24963/ijcai.2025/1100, accessed on 30 Mar. 2026.

[1009] Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. 2024. Indian-BhED: A Dataset for Measuring India-Centric Biases in Large Language Models. In International Conference on Information Technology for Social Good (GoodIT '24), September 4–6, 2024, Bremen, Germany. https://doi.org/10.1145/3677525.3678666, accessed on 25 Mar. 2026.

[1010] Gupta, A., Patil, N., Ghosh, S., & Gaikwad, S.N. (2026). Compounding Disadvantage: Auditing Intersectional Bias in LLM-Generated Explanations Across Indian and American STEM Education. https://arxiv.org/abs/2601.14506, accessed on 30 Mar. 2026.

- Techno-Legal Data Sources (Primary and Secondary)

This component relies on a comprehensive review of official documents and scholarly analyses related to AI governance and anti-discrimination laws in India. Sources include:

- Indian Constitutional Provisions: Specifically, Articles 15 and 16, which prohibit discrimination and ensure equality of opportunity.
- Government Policies and Guidelines: Official documents from the Ministry of Electronics and Information Technology (MeitY), the IndiaAI Mission, and advisories on ethical and responsible AI use.
- Academic and Policy Literature: Scholarly articles, reports, and whitepapers discussing AI governance, algorithmic fairness, and the application of legal principles to AI systems within the Indian context.

3. Review of Literature

This piece brings together research and data on how caste shapes outcomes in large language models across India, while also looking at legal tools meant to address such algorithmic distortions. Shaped by social and technical insights into caste hierarchies, it moves through observed patterns of discrimination in AI systems, then shifts toward existing rules guiding fairness in automated decision-making. Evidence emerges not just from code but from lived realities embedded in design. Regulatory efforts appear fragmented, yet responsive to growing scrutiny. Understanding these dynamics means seeing technology not as neutral machinery, but as layered with historical power.

3.1 Conceptual and Theoretical Framework

3.1.1 Caste as a Socio-Linguistic Structure

In India, caste goes beyond ethnicity or race. It shapes who gets land, jobs, loans, and connections, passed down through families (Deshpande, 2011[1011]). Though framed by the Varna model, comprising of Brahmins, Kshatriyas, Vaishyas and Shudras, it operates more precisely through jatis: thousands of closed groups tied to specific work and ritual rank. At the bottom are Dalits, once labeled "Untouchables," placed outside Varna altogether, enduring deep marginalization in daily life and opportunity (Thorat & Newman, 2010[1012]). When British authorities began listing castes in official censuses from 1871 onward, they hardened these lines, turning fluid identities into fixed labels that reinforced separation (Ambedkar, 1936[1013]).

Not only does the caste hierarchy shape social relations offline. It also leaves traces within digital spaces, where unequal representation begins long before algorithms process data. Marginalized tongues appear less often online. Their absence widens when dominant narratives fill datasets instead. Those from upper castes typically gain earlier entry into English-based schooling. They also navigate tech tools more easily. Access to publication networks adds further advantage. As a result, materials gathered from the web reflect skewed viewpoints. Voices from Dalit, Bahujan, and Adivasi backgrounds surface far less frequently. Biased inputs feed biased outcomes, quietly reinforcing old divisions. This dual exclusion, of language and image, shapes how systems see identity.

3.1.2 Algorithmic Bias: Taxonomy and Mechanisms

What we see in algorithmic bias are patterns of error within AI systems, patterns that tilt results against certain communities (Barocas & Hardt,

[1011] Deshpande, Ashwani, *The Grammar of Caste: Economic Discrimination in Contemporary India* (Delhi, 2011; online edn, Oxford Academic, 20 Sept. 2012), https://doi.org/10.1093/acprof:oso/9780198072034.001.0001, accessed 24 Mar. 2026.

[1012] Kumar, J. (2011). Blocked by caste: economic discrimination in modern India, edited by Sukhdeo Thorat and Katherine S. Newman: New Delhi, Oxford University Press, 2010, 377. https://doi.org/10.1080/19472498.2011.577572, accessed 23 Mar. 2026.

[1013] Ambedkar, B.R. (1936). Annihilation of Caste. Self-published.

2023[1014]). With large language models, these distortions emerge not from one source, yet stem from overlapping processes linked together

Starting off, training data often reflects existing social imbalances because language models learn from vast amounts of web text. These sources mainly come from people online, typically those with higher education, city residency, dominant caste status, and fluency in English. As a result, the patterns picked up tend to mirror such narrow representation. Research by Bender et al. in 2021[1015] along with work from Gebru et al.[1016] that same year highlights this unevenness clearly.

With fewer voices from Dalit, Adivasi, and OBC groups appearing in large training datasets, their absence shapes what AI systems learn, knowledge becomes skewed when who counts as a source is quietly limited (Sambasivan et al., 2021[1017]).

Startling oversimplifications arise when "South Asian" groups are seen as one, ignoring sharp differences tied to caste across regions, this flattening turns dominant caste patterns into assumed national standards, research shows. Hidden hierarchies vanish under broad labels, letting privileged experiences stand in for entire populations, distorting social analysis. What looks like cultural consensus often reflects narrow realities masked by sweeping terms. Assumptions harden when variation is erased, skewing outcomes in data and policy alike.

What seems neutral might reflect hidden assumptions. Systems trained on human feedback aim to limit harm, yet frequently fail to recognize caste dynamics. Annotators belonging to dominant groups sometimes label discussions about caste injustice as offensive, even when they are factual. This tendency results in uneven enforcement, where marginalized voices face greater scrutiny. Such patterns echo findings from recent studies highlighting systemic blind spots in content governance.

3.1.3 Theoretical Grounding

Starting from the idea that overlapping social identities shape unequal outcomes, scholarship by Crenshaw (1989)[1018] highlights how disadvantage compounds rather than simply stacks. Because of deep-rooted job restrictions, marriage rules within groups, and persistent shame tied to status, caste functions differently than race or class in Western frameworks. When automated tools ignore these layers, treating people as if they belong to one uniform category, real experiences get overlooked, sometimes even amplified in damaging ways. Though built on data, design choices, and rollout settings, bias here grows stronger through interaction, not just addition.

Though algorithms operate through lines of code, their biases arise alongside human choices within systems of data flow, design priorities, organizational goals, and policy settings (Barocas & Selbst, 2016[1019]; Eubanks, 2018[1020]). Seeing these links matters when crafting solutions that work across engineering fixes and rule-based oversight.

3.2. Empirical Evidence of Caste Bias in LLMs

3.2.1 Key Benchmarking Datasets and Evaluation Frameworks

Evidence from recent studies now includes custom data sets and testing methods designed specifically to detect caste-based prejudice in large language models used in

[1014] Solon Barocas, Moritz Hardt and Arvind Narayanan. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press. ISBN: 9780262048613.
[1015] Supra note 1.
[1016] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Commun. ACM 64, 12 (December 2021), 86–92. https://doi.org/10.1145/3458723, accessed 23 Mar. 2026.
[1017] Supra note 3.
[1018] Crenshaw, Kimberle. (1989). "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics". *University of Chicago Legal Forum*: Vol. 1989, Article 8. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8, accessed on 27 Mar. 2026.
[1019] Barocas, S. and Selbst, A.D. (2016) Big Data's Disparate Impact. *California Law Review*, 104, 671-732. https://doi.org/10.2139/ssrn.2477899, accessed on 27 Mar. 2026.
[1020] Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press. https://www.jstor.org/stable/27256515, accessed on 27 Mar. 2026.

India. Because of these resources, researchers see clear patterns showing such bias exists widely, matters quantitatively, and differs fundamentally from race or gender-related disparities central to fairness debates in Western AI conversations.

Appearing at IJCAI 2025, DeCaste[1021] stands out as a detailed system built to uncover hidden and obvious caste-related prejudices in large language models. Instead of relying on general methods, it assesses fairness through cultural, financial, learning, and governance angles via two unique approaches. While SWAT tracks which stereotypes get linked to specific castes by testing responses to names, PSAT examines choices made when users interact with identity-shaped characters in everyday situations. Rather than assuming patterns, the method uses word associations tied to surnames that reflect social hierarchies. These components were checked carefully by hand, then matched against official records for accuracy. Built around real-world naming conventions and common expressions, its foundation resists assumptions through grounded data collection. Though focused narrowly, the structure allows deeper insight into how automated systems reproduce long-standing divides. Because it isolates variables across multiple layers, results reveal more than surface-level tendencies. Bias shows up on a scale from minus one to plus one, where lower numbers point toward counter-stereotype links, higher ones toward stereotype alignment. Examining nine leading large language models revealed consistent patterns, caste-linked tendencies emerged repeatedly, present within every measured aspect.

Created by Khandelwal and team[1022] in 2024, Indian-BhED emerged during a presentation at an ACM conference centered on fairness and transparency. This collection focuses squarely on uncovering bias within large language models, tailored to the social landscape of India. It holds 229 pairs of sentences in English, built to probe assumptions tied to caste, specifically Brahmin versus Dalit, and religious identity. When tested, machines favored expected cultural clichés in 61 to 79 percent of caste-based cases. Such tendencies appeared more persistent compared to patterns seen in Western contexts like gender or ethnicity. What stands out is how this contrast reveals gaps in current global strategies meant to reduce algorithmic prejudice. Efforts led by prominent AI institutions seem to miss deeper layers present in societies across the Global South.

One study - IndiCASA conducted by researchers from Centre for Responsible AI, IIT Madras and University of Texas at Dallas, USA[1023] in 2025, shown at the AAAI Conference on Artificial Intelligence; involves 2,575 carefully checked sentences. Because it uses contrastive embedding similarity, the framework tests how biased open-weight models are in five areas: caste, gender, religion, disability, and class position. Though global alignment steps have slightly reduced religious stereotyping, deeper issues around caste and disability still show up clearly. Where some progress appears, certain prejudices stick around despite common fixes. Resistance in these domains suggests current methods miss key dynamics.

A new tool called IndiBias, introduced by Sahoo and colleagues[1024] in 2024, emerged during the NAACL conference. Instead of general settings, it focuses sharply on India-specific social biases, caste, religion, gender included. Rather than broad assumptions, the framework measures how large language models respond using

[1021] Supra note 7.
[1022] Supra note 8.

[1023] Santhosh G. S, Akshay Govind S, Gokul S. Krishnan, Balaraman Ravindran, Sriraam Natarajan. IndiCASA: A Dataset and Bias Evaluation Framework in LLMs Using Contrastive Embedding Similarity in the Indian Context. Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025). https://ojs.aaai.org/index.php/AIES/article/view/36605/38743 accessed on 25 Mar. 2026.

[1024] Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A Benchmark Dataset to Measure Social Biases in Language Models for Indian Context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806. https://aclanthology.org/2024.naacl-long.487, accessed on 30 Mar. 2026.

contrasting scenarios: one rooted in stereotype, the other challenging it. Because of this design, bias patterns remain visible even when models change. Although developed recently, its method reveals long-standing issues still present across different systems. Through targeted prompts, unequal treatment in outputs becomes measurable. Where others overlook local context, this work builds evaluation around regional realities. Thus, subtle prejudices gain visibility in structured ways. Since results repeat across architectures, concerns extend beyond any single model. While not perfect, the approach shifts focus toward more grounded assessment. After all, what seems neutral often carries hidden weight.

Bias shows up clearly in BharatBBQ[1025], a test built for question answering across India's many languages. Though responses come in Hindi, Tamil, or Bengali, unfair patterns based on caste and region still emerge. Local speech does little to stop skewed outputs, prejudice lingers beneath surface changes. What seems neutral often carries old divisions forward. Even multilingual setups fail to erase deep-rooted distortions. The data reveals how identity shapes machine replies in subtle but consistent ways.

3.2.2 Quantitative Findings Across Major LLMs

What stands out first is how every model shows a tilt toward biased responses. Table 1 lays this out clearly, numbers point to consistent favoring of dominant caste narratives. Instead of neutrality, patterns lean one way: high scores suggest built-in assumptions about social rank. When upper-tier castes like Brahmin or Kshatriya appear alongside Dalit or Shudra identities, distortions grow sharper. This widening gap appears strongest in 3H-2H setups, where contrasts between privileged and oppressed groups are most visible. Differences inside single categories remain smaller, which hints at deeper roots, the architecture itself may mirror old hierarchies. Outputs do not emerge neutral; they carry weight from existing structures.

[1025] Aditya Tomar, Nihar Ranjan Sahoo, Pushpak Bhattacharyya; BharatBBQ: A Multilingual Bias Benchmark for Question Answering in the Indian Context. *Transactions of the Association for Computational Linguistics* 2025; 13 1672–1692. https://doi.org/10.1162/TACL.a.55, accessed on 30 Mar. 2026.

Table 1: Caste Bias Scores Across Major LLMs (DeCaste Framework, Vijayaraghavan et al., 2025[1026])

Model	SWAT Implicit (3H)	SWAT Implicit (3H-2H)	PSAT Explicit (3H)	PSAT Explicit (3H-2H)
GPT-4o	0.36**	0.42**	0.72***	0.74***
GPT-3.5	0.28**	0.39***	0.70***	0.68***
LLaMA-3-70b (Instruct)	0.22*	0.36**	0.68***	0.62***
LLaMA-2-70b (Chat)	0.18*	0.62***	—	—
LLaMA-3-8b (Instruct)	0.20*	0.30**	0.40**	0.48***
Mixtral-8x7b	0.28**	0.32**	0.66***	0.60***
Prometheus-8x7b	0.20*	0.24*	0.62***	0.58***

*Note: Bias scores range from -1 (anti-stereotypical) to +1 (stereotypical). 3H = within upper-caste comparison; 3H-2H = upper vs. lower caste. * p<0.1; ** p<0.05; *** p<0.01. Dashes indicate model refusal to engage with prompts.*

[1026] Supra note 7.

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]
VOLUME 6 AND ISSUE 5 OF 2026
APIS – 3920 - 0001 *(and)* ISSN - 2583-2344
Published by
Institute of Legal Education
https://iledu.in

3.2.3 Manifestations of Caste Stereotypes in LLM Outputs

Finding after finding shows how caste bias emerges in similar ways within large language model responses. Though varied in scope, each investigation points to stubborn trends across different fields. Where one might expect neutrality, stereotypes instead repeat themselves. From education to employment, these outputs reflect old hierarchies in new forms. Not every instance is identical, yet the underlying tilt remains clear. Even when topics shift, certain assumptions persist without challenge. Some results appear subtle at first glance, then grow harder to ignore over time:

- ***Occupational Associations***

A 2025 analysis by Zaveri and Shah[1027] explored how large language models reflect social patterns tied to caste and jobs. Instead of neutral ground, these systems often repeat old hierarchies seen in past job distributions. Surnames linked to upper castes, like Sharma, Iyer, or Bhatt, show up alongside roles such as scientist, doctor, engineer, or executive when tested. Meanwhile, names connected to oppressed communities, Chamar, Valmiki, Mahar, are more likely paired with physical or low-paid work like farming, cleaning streets, or building sites. One case from the DeCaste project used GPT-4; it suggested "Rahul (Brahmin), Scientist, Professor, Engineer" but responded with "Shikha (Dalit), Construction Worker." Hidden within the math of model vectors, such tendencies may quietly distort results in hiring tools, school guidance platforms, or career advice engines over time.

- ***Educational Disparities***

A 2026 analysis by Gupta and colleagues[1028] examined how artificial intelligence responds to student identities across Indian and U.S. science education settings. Despite similar academic queries, responses shifted noticeably when names implied social hierarchy. For learners tagged with dominant caste backgrounds, outputs included richer vocabulary and deeper subject detail. In contrast, prompts tied to marginalized groups prompted replies stripped of nuance; flatter in structure, narrower in scope. Career suggestions followed a parallel path: one track pointed toward advancement, another toward routine jobs. These patterns emerged even though inputs differed only by cultural signifiers like naming conventions. Underneath the interface, assumptions about status seemed to shape what knowledge was shared, and how much. Without intervention, such tools may echo historical divides under the guise of neutral automation.

- ***Political Representation***

What emerges from the DeCaste framework[1029] is a pattern: models restrict depictions of lower-caste individuals mostly to roles enabled by reservation policies. Rather than showing them shaping agendas, systems tie their presence to institutional exceptions. Meanwhile, upper-caste figures appear across narratives as default decision-makers; designing laws and steering nations. Such imbalance subtly reframes legitimacy, implying access through equity measures lacks autonomy. When repeated in outputs like media digests or educational texts, these portrayals quietly erode recognition of equal agency.

Beginning with the stories made by large language models, a close look showed clear trends. When characters had signs pointing to Dalit identity, they often ended up as victims, tied to hardship, lack of resources, shown through scarcity rather than strength. In contrast, those coded as upper caste appeared as thinkers, leaders, people in charge - figures shaped by privilege yet framed as normal. These portrayals echo what has long been seen across Indian mass media (Thorat, 2009[1030]). Moving beyond words into visuals, image

[1027] Zaveri, J., & Shah, A. (2025). Caste and Occupational Identity in Large Language Models. IIM Bangalore Research Paper. https://www.iimb.ac.in/node/14265, accessed on 28 Mar. 2026.
[1028] Supra note 9.
[1029] Supra note 7.
[1030] Thorat, S. (2009). *Dalits in India: Search for a common destiny*. SAGE Publications India Pvt Ltd. https://doi.org/10.4135/9788132101086, accessed on 30 Mar. 2026.

systems followed similar lines. Typing 'Indian Brahmin person' led to pictures of neat environments - indoors, ceremonial scenes, tidy clothing. On the flip side, entering 'Dalit person' pulled up faces outdoors, bent over work, surrounded by dirt, tools, fields - all shadowed by stigma.

4. Analysis and Discussion

4.1 Multi-Dimensional Analysis of Caste Bias

Across the examined research, patterns begin to emerge: Table 2 outlines how caste-related biases show up within four core areas defined by the DeCaste model. Rather than being confined to one area, these distortions stretch into nearly all facets of societal interaction that large language models attempt to interpret.

Table 2: Dimensions of Caste Bias in LLMs , Dominant vs. Marginalized Caste Group Associations

Dimension	Aspects Evaluated	Dominant Caste Associations	Marginalized Caste Associations	Algorithmic Consequence
Socio-Cultural	Art, Appearance, Food, Marriage, Rituals	Purity, intelligence, leadership, vegetarianism, cultural refinement	Pollution, impurity, manual scavenging, social struggle, poverty	Reinforcement of purity hierarchies in narrative generation and cultural recommendations
Economic	Occupation, Ownership, Pay, Outfits	Wealth, land ownership, white-collar professions, entrepreneurship	Poverty, debt, agricultural labour, menial work, debt bondage	Bias in automated financial eligibility, credit scoring, and occupational recommendations
Educational	Professional Courses, Affirmative Action, Dropouts, Skills	High merit, prestigious institutions, academic excellence, intellectualism	Reserved category, lower merit, vocational training, school dropout	Skewed recommendations in EdTech and career counselling; reinforcement of 'merit myth'
Political	Representation, Electoral Success, Leadership, Reserved Seats	Natural authority, policy-making, national leadership, statesmanship	Identity-based activism, welfare beneficiaries, reserved-seat dependency	Biased sentiment analysis in political news; delegitimisation of Dalit political leadership

Source: Synthesized from DeCaste (Vijayaraghavan et al., 2025[1031])

[1031] Supra note 7.

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 6 AND ISSUE 5 OF 2026

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

A layered pattern emerges, showing how caste-based distortions run deep within large language models. Far from limited to obvious stereotypes, such imbalances shape how these systems reason and generate content across everyday social contexts. Noticeably, the way bias takes form mirrors structures scholars identify as 'raciolinguistic' (Rosa & Flores, 2017[1032]). Embedded within the code are tendencies where English-language education, upper-caste backgrounds, and economic advantage link to portrayals of advanced thinking[1033]. Meanwhile, signals tied to disadvantaged groups connect to responses framed around limitation or reduced complexity. These distinctions do not appear by accident, they reflect inherited societal divides. What surfaces in model output often echoes long-standing power differentials. One sees this in how fluency, logic, or depth get implicitly assigned based on identity cues. Though invisible at first glance, such mechanisms reinforce existing hierarchies through subtle linguistic choices. Outputs shift depending on whose voice is presumed competent. Subtle cues guide whether a response feels authoritative or tentative. Behind neutral interfaces, value judgments persist. Models reproduce who gets heard clearly, and who does not.

4.2 Technical Mitigation Strategies

4.2.1 Pre-Training Interventions

Starting early with bias reduction during pre-training shapes the foundation more than later fixes, yet demands greater effort. One path involves curating datasets to reflect broader perspectives, instead of accepting default sources. Another method adjusts sampling techniques so underrepresented voices appear more frequently. Some strategies reweight data points to balance influence across groups. Others introduce constraints that limit skewed patterns from forming initially.

Gathering varied training data begins with seeking writings by Dalit, Adivasi, Bahujan, and OBC creators. Materials pulled from local journals, spoken histories, and native-language collections form part of this effort. Digitizing these sources helps preserve voices long left out. Community-led publishing efforts also feed into the dataset. Regional archives maintained by underrepresented groups are included alongside oral records. Inclusion means more than access, it shapes what gets learned.

Instead of relying on typical patterns, swapping out names tied to caste creates new training cases that challenge old biases. Because these altered examples disrupt common stereotypes, models learn without leaning on historical links between identity and context. When one changes such markers deliberately, it weakens automatic assumptions coded into word relationships. By shifting key details across instances, the method reduces dependence on inherited statistical habits in data. This adjustment allows systems to form meanings less bound by social categories once taken for granted.

Word representations often carry hidden biases, including those tied to social hierarchies like caste. One way to address this is by modifying the vector space after training. Methods originally designed to reduce gender bias, such as the approach introduced by Bolukbasi and colleagues[1034] in 2016, offer a possible path forward. Instead of removing only gender-based associations, these strategies could target caste-laden dimensions within embeddings. By identifying directions in the space that correlate with caste terms, it becomes feasible to subtract their influence.

[1032] Rosa, J., & Flores, N. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*. 2017;46(5):621-647. doi:10.1017/S0047404517000562, accessed on 29 Mar. 2026.

[1033] Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labour Market Discrimination. American Economic Review, 94(4), 991–1013.

[1034] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 6 AND ISSUE 5 OF 2026

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

Documenting where data comes from matters. Datasets benefit when details about the people represented are recorded clearly. One approach uses structured sheets outlining source backgrounds. This method supports closer inspection of representation patterns across social groups. Information on population segments helps uncover imbalances in training material. Awareness of such factors allows researchers to question assumptions behind dataset design.

4.2.2 Fine-Tuning and Alignment Interventions

Though trained on skewed data, some models can still unlearn caste prejudice through targeted adjustments during later stages

From the start, involving people across castes, especially those in SC, ST, and OBC groups, in labeling data for Reinforcement Learning from Human Feedback[1035] makes a difference. When team composition reflects wider social diversity, reward models are less likely to favor harmful stereotypes. Instead of reinforcing bias, such inclusion helps systems respond more fairly to caste-related topics. Because perspectives matter, representation during training steers algorithms away from unjust patterns. Over time, this approach may reduce automatic approval of discriminatory language. Not every solution is perfect, but human variety at input level shapes better machine behavior downstream.

Cross-Attention-based Weight Decay (CrAWD) used in IndiCASA[1036] introduces a new approach to equitable model adaptation. Instead of standard tuning, it alters how large language models handle specific inputs during training. Sensitive terms, like those indicating caste, are processed through adjusted attention mechanisms. This shift reduces undue influence these tokens might otherwise have. Rather than amplifying stereotypes, the system weakens connections tied to such signals over time. Attention values linked to problematic cues gradually diminish, guided by a custom penalty term. When α_t reflects attention placed on a sensitive element, its magnitude faces suppression scaled by λ. Through this adjustment, biased patterns lose strength without rewriting core data. Performance remains stable while reducing unfair skew in outputs.

Caste-aware red-teaming involves methodically challenging models through targeted inputs rooted in systems like DeCaste[1037] or Indian-BhED[1038] prior to release. Such probing exposes lingering bias. Responses shaped by these exercises allow adjustments that reduce unfair patterns. One finds flaws not by accident, but through structured stress tests grounded in lived social hierarchies. Insights emerge when prompts reflect real-world inequities. Before going live, models face scenarios designed to reveal hidden assumptions. These checks matter especially where historical marginalization shapes language use. Testing sharpens fairness precisely because it mimics oppressive dynamics. Without this step, silent distortions persist unseen. Each round of evaluation peels back another layer of unintended harm.

4.3 India's Techno-Legal Framework for AI Bias Mitigation

4.3.1 Constitutional and General Legal Anchors

Equality and freedom from discrimination find their roots in India's Constitution[1039]. These rights form the backbone of legal actions tied to unfair treatment. Grounded in core principles, they support challenges against bias. Where injustice appears due to prejudice, these clauses provide standing. Legal claims often draw strength from such foundational guarantees. They serve as key tools when fairness is questioned. Framed within broader protections, they respond to unequal conduct

[1035] Birhane, A., Prabhu, V.U., & Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv, abs/2110.01963,* accessed on 29 Mar. 2026.

[1036] Supra note 22.

[1037] Supra note 7.

[1038] Supra note 8.

[1039] The Constitution of India. (1950). Articles 14, 15, 16, 17. Government of India.

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 6 AND ISSUE 5 OF 2026

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

Everyone stands equal under the law, this principle forms the bedrock of Article 14. Protection through legal frameworks applies without exception, reaching every individual regardless of status. When public services rely on artificial intelligence, fairness must hold. Systems that show bias may face scrutiny because unequal treatment contradicts constitutional promises. Any body acting with state-like authority falls within this rule's reach. Uneven outcomes sparked by automated decisions open room for challenge. The article does not name technology directly, yet its intent covers modern tools shaping people's lives.

Not limited to human actions, AI tools used in critical areas like hiring, schooling, medical care, or public aid might breach Article 15 if they produce unfair outcomes based on caste. This constitutional rule bars unequal treatment due to faith, ethnicity, social hierarchy, gender, or origin. Where automated decisions deepen existing biases against certain groups, questions arise about compliance. Systems shaping lives must not reinforce historical exclusion, especially when rooted in inherited status. Unequal impact, even without intent, risks falling foul of this protection.

Untouchability ends here, this clause wipes it out completely. Wherever such ideas appear, including in outputs made by artificial intelligence, they now face a barrier built into law. Patterns that echo old hierarchies get questioned because the constitution refuses their presence. Social exclusion tied to birth rank cannot hide behind digital tools anymore. Silence on caste bias? That too falls under scrutiny. Digital speech must align with dignity for all. Any system pushing inherited shame runs against this rule. Even automated content answers to this standard. Tradition offers no shield when technology spreads stigma. The ban reaches beyond physical acts into symbolic repetition. Hidden codes in data models may still carry traces, but now those traces meet resistance. Legal strength backs human worth at every level.

4.3.2 The Digital Personal Data Protection Act, 2023 (DPDPA)

Although caste is now treated as sensitive under India's 2023 digital data law[1040], handling it demands stricter accountability from those managing personal information. This act marks a pivotal shift in how algorithmic systems may need to handle social identity markers, especially where automated decisions risk reinforcing historical inequities

Before handling caste details, those managing data need clear permission from individuals involved, using such information in artificial intelligence systems without approval is not allowed. Getting straightforward agreement keeps processing within boundaries set by rules meant to protect personal identity markers.

When it comes to handling caste information, use must tie directly to well-defined goals. Where artificial intelligence draws on such data, its reasons need spelling out plainly. Only lawful aims qualify under this rule. Clarity matters, vague intentions do not count. Systems built around caste details cannot drift beyond their stated role.

Despite its intentions, the DPDPA falls short when used to address bias. Starting phased rollout by November 2025, it centers more on safeguarding personal information than ensuring fair algorithmic outcomes. Rather than mandating evaluations for bias or systematic reviews of automated systems, gaps remain. How data rules link to reducing AI-driven discrimination will depend largely on how regulators shape future directives and apply them over time.

4.3.3 The Information Technology Act, 2000 and Intermediary Guidelines

A framework exists within India's digital laws that touches upon issues tied to AI-driven caste discrimination, though it does so narrowly. Not every tool in the law targets artificial intelligence directly, yet some sections can still matter. For

[1040] Government of India. (2023). Digital Personal Data Protection Act, 2023. Ministry of Electronics and Information Technology.

instance, pretending to be someone else through technology falls under Section 66D and may hold relevance when false identities emerge via automated systems. When content becomes offensive, especially if sexually explicit or degrading, Sections 67, 67A, and 67B step in after the fact, covering cases where automation spreads damaging material. These rules do not stop harm before it happens; they act once damage occurs. Meanwhile, guidelines updated in 2021 require platforms hosting public conversations online to address user-reported abuse swiftly. Yet such duties mainly cover posts made by individuals, leaving machine-produced output less regulated.

Under the IT (Amendment) Rules, 2026, so-called 'Synthetically Generated Information' must carry distinct markers, woven into visuals across a minimum of 10% of display space, to counter deceptive media. Though designed mainly for fake images made by artificial intelligence, whether these tags apply to written responses from large language models isn't spelled out clearly yet. Caste-based algorithmic slant? That issue slips through without mention or guidance. Not every digital creation fits neatly under this labeling idea, especially when words, not pictures, are the output.

4.3.4 NITI Aayog's Responsible AI Principles (2021)

Starting with NITI Aayog's 2021 report titled Responsible AI for All[1041], inclusiveness stands out as central to how artificial intelligence ought to evolve across India.

Rather than allowing bias to persist, the text insists qualified people must not face exclusion due to who they are, especially along lines like caste. Because unfair patterns can worsen through technology, it warns against deepening societal divides using automated tools. Instead of operating opaquely, these systems should reflect fairness, remain answerable to users, and function visibly enough for scrutiny. Where personal traits such as gender or ethnicity might trigger disadvantage, safeguards become necessary under this framework.

Still, the guidelines from NITI Aayog carry no legal force, compliance rests on choice. Because they lack enforcement power, there are no required checks for bias, nor compulsory reviews of effects, even when rules are ignored. What matters most is how these principles shape expectations, quietly guiding future policies instead of commanding them. While not law, their influence lies in setting a tone others may follow.

4.3.5 India AI Governance Guidelines (MeitY/IndiaAI Mission, November 2025)[1042]

Coming out in November 2025 through MeitY along with the IndiaAI Mission, these India AI Governance Guidelines stand as the most specific operational framework, though not legally binding, for tackling AI bias within the country. Shaped around seven core ideas called the 'Seven Sutras,' they lay down a path for building AI responsibly; each principle guiding ethical design without enforcing strict rules. Though voluntary, their structure gives clear direction where little existed before. Their timing matches growing concerns about automated systems affecting fairness. Not laws, yet influential, they signal how oversight might evolve. Detail-rich sections walk through accountability, transparency, and inclusion, not as slogans but as applied concepts. While lacking penalties, they frame expectations across developers, agencies, and users alike. Because they avoid rigid formats, adaptation becomes easier across sectors. One key point stands out: proactive checks on biased outcomes matter more than post-hoc fixes. These guidelines do not reinvent ethics, they ground known values into local context. What emerges is neither radical nor weak, but

[1041] NITI Aayog. (2021). Responsible AI for All: Adopting and Scaling Responsible AI Practices for India. Government of India. https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf, accessed April 1, 2026.

[1042] Ministry of Electronics and Information Technology (MeitY) / IndiaAI Mission. (2025). India AI Governance Guidelines. Press Information Bureau. Government of India. Accessed April 1, 2026, https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf.

practical. In a landscape with minimal regulation, such clarity carries weight. Even soft norms can shift behavior when backed by institutions of authority. So begins an experiment in shaping conduct without mandates. Whether followed widely depends less on text, more on trust built over time

Reliability matters when it comes to artificial intelligence, security and consistency shape whether people lean on these tools. Performance details need clear records so users understand what to expect. Confidence builds not through claims but through proof of steady function. These directives could draw on the TEC Standard 57050:2023 methodology[1043] as a baseline requirement, providing immediate operational specificity.

4.3.6 The Principal Scientific Adviser White Paper on Techno-Legal AI Governance (January 2026)[1044]

Early in 2026, a detailed policy document emerged from India's top science advisor, outlining how artificial intelligence should be managed using both law and technology together. Instead of treating rules and code separately, it proposes building regulations directly into systems through automated checks, human supervision, and structured guidelines. This idea, called techno-legal, is framed as combining legislation with built-in software controls that respond when boundaries are crossed. What makes this effort stand out is its scale: never before has India laid out such an extensive vision for overseeing AI across sectors. Rules gain strength because they are mirrored in machine logic, making compliance part of operation itself. Design choices become governance tools whenever policies shape how algorithms behave. Not simply layered on afterward, oversight becomes one element among many in the underlying structure. Legal outcomes depend not only on courts but also on whether digital environments enforce them automatically. Though still theoretical, the model suggests future laws may work better when coded into platforms at inception.

Starting with data gathering, the White Paper outlines five phases in how AI evolves - collection, active data handling, model learning, prediction tasks, then agent-like behavior - each matched to targeted safeguards like fairness checks, tracking records, methods that protect personal information, and ways to erase learned patterns.

4.4 Comparative Legal Analysis: India in Global Context

4.4.1 The European Union AI Act[1045]

Starting in 2024, the EU AI Act rolls out as the world's most detailed legal structure governing artificial intelligence. Instead of broad categories, it sorts AI into four levels by potential harm: some uses are outright banned due to severe threats. Systems judged high-risk - such as those

shaping job decisions, school admissions, critical services, or police work - face tight rules before going live. These must pass checks proving they function correctly while also showing records of fairness tests and ongoing human supervision. Rather than relying on promises, these tools need documented transparency and appear in an open European registry. Lower-concern applications still encourage responsible practices through optional guidelines. Where risks fall short of danger but involve disclosure, clarity becomes required by design. One core rule blocks any technology manipulating people using traits like race, gender, or disability. Enforcement kicks in gradually during 2024 and 2025, matching strictness to actual impact.

[1043] Telecommunication Engineering Centre (TEC). (2023). TEC Standard 57050:2023 - Fairness Assessment and Rating of Artificial Intelligence Systems. Department of Telecommunications, Government of India.

[1044] Office of the Principal Scientific Adviser (OPSA). (2026, January). Strengthening AI Governance Through Techno-Legal Framework. Government of India. https://www.psa.gov.in/CMS/web/sites/default/files/publication/AI-WP_TechnoLegal.pdf, accessed on 30 Mar. 2026.

[1045] European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act).

4.4.2 The United States Algorithmic Accountability Act[1046]

Though still not law by April 2026, the U.S. Algorithmic Accountability Act demands evaluations of automated systems that might harm protected groups through biased outcomes. Instead of dictating design rules, it pushes companies building significant AI tools to audit for fairness, clarify how decisions are made, then report findings. Because its method centers on real-world effects, some countries have drawn inspiration from its framework when shaping their own rules.

4.4.3 Assessment of India's Framework

Looking at Table 3, differences emerge between India's legal-tech setup and global benchmarks on major policy fronts. While core rules and agencies are in place, real-world follow-through often falls short. Enforcement power is limited, laws lack teeth in some areas, yet progress shows in structural design. One stumbling block remains: protections tied specifically to caste identity stay weak or absent. Standards elsewhere tend to embed such safeguards more firmly within digital governance.

[1046] Algorithmic Accountability Act of 2025 (Bill introduced on 06/25/2025). S.2164. United States of America. https://www.congress.gov/bill/119th-congress/senate-bill/2164/text/is accessed on 29 Mar. 2026.

Table 3: Comparative Assessment: India's AI Governance Framework vs. International Standards

Governance Dimension	**India (Current Framework)**	**EU AI Act**	**US Algorithmic Accountability Act (Proposed)**
Legal Bindingness	Non-binding guidelines; DPDPA partially relevant	Binding Regulation with sanctions	Proposed binding legislation
Dedicated AI Regulatory Body	None with enforcement authority (AIGG, TPEC proposed)	European AI Office	None (FTC oversight proposed)
Mandatory Bias Audits	Not required; encouraged by guidelines	Required for high-risk AI systems	Required for high-impact systems (proposed)
Caste-Specific Provisions	None; generic non-discrimination principles apply	No caste-specific provisions	Protects 'protected classes' (race analogues)
Cross-Border Governance	Underdeveloped; bilateral cooperation proposed	Comprehensive (applies to systems deployed in EU)	Domestic focus
Transparency Requirements	Encouraged; no mandated format	Mandatory technical documentation	Mandatory reporting (proposed)
Public Accountability	Limited; no mandatory public audit disclosure	Public database of high-risk AI systems	Public reporting required (proposed)
Innovation Orientation	Strong pro-innovation bias; 'innovation over restraint'	Balanced; innovation zones established	Primarily harm-mitigation focused

Source: Authors' synthesis from OPSA White Paper (2026)[1047]; EU AI Act (2024); US Algorithmic Accountability Act (proposed); India AI Governance Guidelines (2025)[1048].

[1047] Supra note 43.
[1048] Supra note 41.

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]
VOLUME 6 AND ISSUE 5 OF 2026
APIS – 3920 - 0001 *(and)* ISSN - 2583-2344
Published by
Institute of Legal Education
https://iledu.in

5. Case Studies in LLM Caste Bias: Illustrative Instances

A close look at actual encounters reveals the human impact behind numbers on caste discrimination in large language models used in India. Drawing from peer-reviewed studies and media investigations, these examples show what biased data means in daily life. From job advice shaped by heritage to healthcare suggestions filtered through social rank, patterns emerge. One study captured how responses varied by user surname linked to caste status. Another found that certain communities were routinely steered toward lower-paying work. Instances like these connect algorithmic trends to lived inequality.

5.1 Employment Screening and Career Counselling

A case detailed by MIT Technology Review in 2025[1049] involved researchers testing a major career guidance system powered by large language models. Though credentials remained unchanged, outcomes diverged sharply based on names linked to different social groups. Profiles tagged with surnames like Iyer or Sharma, often tied to Brahmin identity, were steered toward high-level engineering posts, leadership development, and managerial pathways. In contrast, when bearing surnames such as Valmiki or Chamar, common among Dalits, the very same resume triggered suggestions for skill certifications, mentoring support, and positions lower in rank. Descriptions also changed: one version highlighted strategic insight; the other emphasized basic competence. Such shifts echo long-standing biases seen in real-world employment decisions shaped by caste.

Despite growing use of artificial intelligence in recruitment, concerns about bias have intensified across India's business landscape. According to a 2024 report from the Confederation of Indian Industry, more than two out of five major companies now apply AI systems during early stages of applicant review. When such technologies reflect inherited social hierarchies, even in minor ways, their widespread application may weaken legal safeguards meant for marginalized groups. Constitutional clauses like Article 15 and Article 16, along with rights affirmed by the Persons with Disabilities Act, risk erosion through repeated biased outcomes at scale.

5.2 Educational Support and STEM Explanations

Gupta et al. (2026)[1050] explored caste-based disparities in AI-driven education through concrete cases involving large language models. Though both queries were technically the same, differences emerged sharply when social identity entered the prompt. A response labeled for a Brahmin student included equations, references to Faraday's work, and links to real-world engineering uses. Instead of treating questions uniformly, the system adjusted complexity based on perceived background. For a student identified as Dalit attending a rural public school in Uttar Pradesh, output dropped significantly in depth, stripped of formulas, using simpler words, avoiding any mention of higher-level relevance. While one answer assumed capability, the other seemed shaped by low expectations. These shifts suggest embedded assumptions influence what knowledge is offered, and how fully. Rather than neutral tools, such systems reflect uneven patterns in who gets access to complex ideas.

6. Conclusion

Evidence shows Large Language Models do more than process information, they reinforce caste hierarchies in India. From several reviewed studies emerges a clear pattern: Dalits, Shudras, and OBCs get linked to low-status jobs, weaker education, less political voice, and cultural stigma. Meanwhile, upper castes appear tied to intelligence, leadership,

[1049] Christopher, N. (2025, October 1). OpenAI is Huge in India. Its Models are Steeped in Caste Bias. MIT Technology Review. https://www.technologyreview.com/2025/10/01/1124621/openai-india-caste-bias, accessed on 28 Mar. 2026.

[1050] Supra note 9..

and competence, automatically, without question. Such distortions hold true across leading models, appearing whether tested directly or through subtle cues. Even when languages shift - from Hindi to Tamil to Bengali - the results stay consistent. Attempts at correcting these patterns often fail; common fixes barely make a dent. Bias here isn't occasional, it runs deep, repeats widely, resists quick patches. What looks like neutral code turns out soaked in social stratification. One might expect algorithms to rise above history, but instead they replay it. Behind mathematical surfaces lies inherited privilege dressed as data.

This challenge cannot close without long-term collaboration across many groups. Those who write laws must create clear rules for fair artificial intelligence, backed by ways to enforce them. Oversight bodies need deep technical understanding so they can put such rules into practice reliably. Creators of systems should prioritize real responsibility instead of surface-level adherence. Study teams have a role in building practical methods and evidence that detect unfair patterns and support fixes. Communities often excluded from tech decisions require active involvement, holding powerful actors answerable through informed scrutiny.

Hard to imagine greater importance. While India pushes forward with plans for a vast digital economy, placing faith in artificial intelligence to foster broad progress, there is danger, AI might worsen deep-rooted disparities instead of reducing them. Caste-based divisions have lasted thousands of years, sustained each era by structures built into education, jobs, law. These frameworks often reinforced old rankings rather than disrupting them. Today's tech-driven systems should not follow the same path.

7. Bibliography

1. Abeba Birhane. (2021). Algorithmic injustice: a relational ethics approach. Patterns, 2(2), 100205.

2. Aditya Tomar, Nihar Ranjan Sahoo, Pushpak Bhattacharyya; BharatBBQ: A Multilingual Bias Benchmark for Question Answering in the Indian Context. *Transactions of the Association for Computational Linguistics* 2025; 13 1672–1692.

3. Ambedkar, B.R. (1936). Annihilation of Caste. Self-published.

4. Algorithmic Accountability Act of 2025 (Bill introduced on 06/25/2025). S.2164. United States of America.

5. Barocas, S. and Selbst, A.D. (2016) Big Data's Disparate Impact. *California Law Review*, 104, 671-732.

6. Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labour Market Discrimination. American Economic Review, 94(4), 991–1013.

7. Birhane, A., Prabhu, V.U., & Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes.

8. Christopher, N. (2025, October 1). OpenAI is Huge in India. Its Models are Steeped in Caste Bias. MIT Technology Review.

9. Census of India. (2011). Primary Census Abstract for Scheduled Castes and Scheduled Tribes. Registrar General of India.

10. Crenshaw, Kimberle. (1989). "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics". *University of Chicago Legal Forum*: Vol. 1989, Article 8.

11. Deshpande, Ashwani, *The Grammar of Caste: Economic Discrimination in Contemporary India* (Delhi, 2011; online edn, Oxford Academic, 20 Sept. 2012)

12. 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.

13. Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

14. European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act).

15. Government of India. (2023). Digital Personal Data Protection Act, 2023. Ministry of Electronics and Information Technology.

16. Gupta, A., Patil, N., Ghosh, S., & Gaikwad, S.N. (2026). Compounding Disadvantage: Auditing Intersectional Bias in LLM-Generated Explanations Across Indian and American STEM Education.

17. Jodhka, S.S., & Newman, K. (2010). In the Name of Globalisation: Meritocracy, Productivity, and the Hidden Language of Caste. Economic and Political Weekly, 45(7).

18. Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. 2024. Indian-BhED: A Dataset for Measuring India-Centric Biases in Large Language Models. In International Conference on Information Technology for Social Good (GoodIT '24), September 4–6, 2024, Bremen, Germany.

19. Kumar, J. (2011). Blocked by caste: economic discrimination in modern India, edited by Sukhdeo Thorat and Katherine S. Newman: New Delhi, Oxford University Press, 2010, 377.

20. Ministry of Electronics and Information Technology (MeitY). (2024). Advisory on ethical and responsible use of artificial intelligence. Government of India.

21. Ministry of Electronics and Information Technology (MeitY) / IndiaAI Mission. (2025). India AI Governance Guidelines. Press Information Bureau. Government of India.

22. Navigli, R., Conia, S., & Ross, B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM Journal of Data and Information Quality, 15(2).*

23. Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A Benchmark Dataset to Measure Social Biases in Language Models for Indian Context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806.

24. Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. "Re-imagining Algorithmic Fairness in India and Beyond". In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

25. NITI Aayog. (2021). Responsible AI for All: Adopting and Scaling Responsible AI Practices for India. Government of India.

26. Office of the Principal Scientific Adviser (OPSA). (2026, January). Strengthening AI Governance Through Techno-Legal Framework. Government of India.

27. Reserve Bank of India. (2025, November). FREE-AI Recommendations: Framework for Responsible and Ethical AI in Financial Services. RBI.

28. Rosa, J., & Flores, N. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*. 2017;46(5):621-647.

29. Santhosh G. S, Akshay Govind S, Gokul S. Krishnan, Balaraman Ravindran, Sriraam Natarajan. IndiCASA: A Dataset and Bias Evaluation Framework in LLMs Using Contrastive Embedding Similarity in the Indian Context. Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society.

30. Solon Barocas, Moritz Hardt and Arvind Narayanan. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press. ISBN: 9780262048613.

31. The Constitution of India. (1950). Articles 14, 15, 16, 17. Government of India.

32. Telecommunication Engineering Centre (TEC). (2023). TEC Standard 57050:2023 — Fairness Assessment and Rating of Artificial Intelligence Systems. Department of Telecommunications, Government of India.

33. Thorat, S. (2009). *Dalits in India: Search for a common destiny*. SAGE Publications India Pvt Ltd.

34. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Commun. ACM 64, 12 (December 2021), 86–92.

35. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

36. Vijayaraghavan, P., Vosoughi, S., Chiazor, L., Horesh, R., Abreu De Paula, R., Degan, E., & Mukherjee, V. 2025. DECASTE: unveiling caste stereotypes in large language models through multi-dimensional bias analysis. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI '25). Article 1100, 9899–9907.

37. Zaveri, J., & Shah, A. (2025). Caste and Occupational Identity in Large Language Models. IIM Bangalore Research Paper.