

Decoding Green Justice: An AI-Assisted Exploration of Indian Environmental Court Rulings over Five Decades

Patrick Behrer¹, Shareen Joshi², Olexiy Kyrychenko³, Viknesh Nagarathinam⁴, Peter Neis⁵ and Shashank Singh⁶

¹World Bank, Washington DC, USA

²Georgetown University, Washington DC, USA

³Radboud University, the Netherlands

⁴Georgetown University, Washington DC, USA

⁵Université Clermont Auvergne, CNRS, IRD, CERDI, France

⁶The University of Chicago, Illinois, USA

Abstract

This study demonstrates the potential of large language models (LLMs) to analyze environmental court rulings from India. Using a novel dataset of 12,615 environmental court orders spanning three decades, we evaluate the performance of two LLMs - GPT-4 API and Claude 3.5 Sonnet - in coding and interpreting judicial decisions. The LLMs are tasked with identifying pro-environmental rulings and extracting key case attributes, with their performance benchmarked against human coders who analyzed 1,910 rulings. Both models achieve approximately 70% accuracy compared to human coding, with the GPT-4 API showing slightly better performance in various sub-samples. These findings suggest promising applications for AI to improve access to and analysis of legal data, particularly in jurisdictions where administrative records lack standardization.

Keywords

Environmental Law, Large-Language Models, Argument Mining, India

1. Introduction

Environmental courts issue thousands of complex rulings, collectively shaping policy and regulatory frameworks across jurisdictions [1]. This volume creates an analytical paradox: The judicial decisions most critical to environmental outcomes are too numerous and complex for systematic evaluation, leaving crucial patterns in environmental jurisprudence largely hidden from researchers and policy-makers. This issue is particularly acute in India, where the judiciary has emerged as a global leader in environmental governance [2, 3, 4] but empirical analysis of decisions has been quite limited [5, 6] until recently [7, 8].

The analysis of environmental rulings faces some fundamental limitations. Manual review of thousands of unstructured legal documents is cumbersome and requires specialized expertise [9]. Coding a large number of documents systematically and consistently can thus be prohibitively expensive. Recent advances in Large Language Models (LLMs) thus offer a promising solution, demonstrating strong capabilities in the analysis of complex legal texts [10, 11, 12, 13, 14, 15].

This study examines whether LLMs perform as well as human experts in categorizing rulings as environmental and assessing whether judicial decisions produce positive environmental results in the context of India. Assessment of the feasibility of using LLMs to complete these tasks is important to

Proceedings of the First Argument Mining and Empirical Legal Research Workshop (AMELR 2025), June 20, 2025, Chicago, United States. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

✉ abehrer@worldbank.org (P. Behrer); sj244@georgetown.edu (S. Joshi); olexiy.kyrychenko@ru.nl (O. Kyrychenko); viknesh.n91@gmail.com (V. Nagarathinam); peter.neis@uca.fr (P. Neis); shashanksinghss09@gmail.com (S. Singh)

🆔 0000-0001-6946-7213 (P. Behrer); 0000-0001-5693-7140 (S. Joshi); 0000-0002-3557-7892 (O. Kyrychenko); 0009-0003-3126-5589 (V. Nagarathinam); 0009-0008-4358-4110 (P. Neis); 0009-0003-0692-3850 (S. Singh)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

determine whether they can be used to expand the quantitative analysis of environmental jurisprudence to inform policy and improve access to environmental justice.

We examine a novel data set of 12,615 environmental court rulings from India spanning three decades, evaluating two state-of-the-art LLMs (GPT-4 and Claude 3.5 Sonnet) against human expert coding of 1,910 rulings. Our central task, determining whether a judicial decision is "pro-environment", is a complex judgment requiring understanding of legal reasoning, environmental science, and implementation realities.

Our work makes three primary contributions. First, we develop and validate a methodology for AI-assisted environmental law analysis that achieves approximately 70% agreement with human experts, which is comparable to studies of the US Supreme Court[10]. Second, we create the first comprehensive AI-annotated dataset of 12,615 Indian environmental rulings, providing a valuable resource for legal informatics research. Given India's pioneering role in environmental jurisprudence, this data set enables the analysis of the evolution and impact of judicial environmental protection. Third, we identify systematic differences between AI and human environmental impact assessments, revealing insights about AI capabilities and the complexities of evaluating judicial effectiveness. These differences highlight the gap between formal legal interventions and perceived real-world impact, which is crucial for environmental policy.

2. Data

Our analysis begins with India's three foundational environmental acts: the Water (Prevention and Control of Pollution) Act 1974, the Air (Prevention and Control of Pollution) Act 1981, and the Environment (Protection) Act 1986.¹ We conducted a comprehensive search of the Indian Kanoon.org database, identifying 2,996 judicial rulings that explicitly cited at least one of these acts² To ensure complete coverage, we systematically expanded our data set by analyzing all additional legislative acts cited within this initial corpus, identifying 23 additional environmental statutes frequently referenced in environmental litigation. The most cited acts in our data are presented in Table 1.

Table 1

Acts Cited in Our Database (Acts cited at least 300 times)

| Act | Number of rulings citing Act |
|---|------------------------------|
| Code of Criminal Procedure, 1973 | 3499 |
| Wildlife (protection) Act, 1972 | 2566 |
| Code of Civil procedure, 1908 | 2219 |
| Environment (protection) Act, 1986 | 1667 |
| Water (prevention and control of pollution) Act, 1974 | 1547 |
| Air (prevention and control of pollution) Act, 1981 | 1374 |
| Indian Penal Code, 1860 | 1304 |
| Article 226 in Constitution of India | 1150 |
| Forest (conservation) Act, 1980 | 1059 |
| National Green Tribunal Act, 2010 | 892 |
| Article 21 in Constitution of India | 667 |
| Indian Forest Act, 1927 | 495 |

Our final dataset encompasses all judicial rulings from 1974 onward citing any identified environmental statute, resulting in 12,615 court rulings spanning through 2024. The raw data consisted of

¹This identification was based on extensive desk research, personal interviews with environmental law experts, and consultation of leading environmental law textbooks ([5]). Previous research has argued that these acts are the main legislative tools for environmental protection in India ([6]).

²IndianKanoon.org was selected because it provides free access to a comprehensive database of Indian court judgments and has been widely used in academic research on the Indian legal system [17].

unstructured text documents with varying formats depending on the court and the time period. Each case document contained the full text of the judicial ruling, including case details, facts, legal arguments, and final orders. We developed a systematic processing pipeline to extract key structured information from these documents, including official case numbers and court identifiers; Names of petitioners, respondents, and presiding judges; Case classification details (civil vs. criminal, judgment vs. order); Geographic jurisdiction and relevant locations and citations to environmental statutes and precedents. Document lengths range from brief procedural orders to comprehensive judgments exceeding 50,000 words, with a median of 917 words and mean of 2,614 words.

Table 2 presents the distribution of rulings at different levels of the Indian judiciary system. Most of the rulings (69%) originated in the High Courts, which serve as the primary forums for environmental litigation in the Indian federal system. The National Green Tribunal (NGT), established in 2010 as a specialized environmental court, represents 23.1% of rulings despite its relatively recent creation. The Supreme Court of India, as the apex court, contributed 3.3% of the rulings, usually involving appeals or matters of national importance.

Table 2
Distribution of Rulings by Court Type

| Court Type | Number of Cases | Percentage |
|-------------------------|-----------------|------------|
| High Courts | 8,706 | 69.0% |
| National Green Tribunal | 2,925 | 23.1% |
| Supreme Court of India | 415 | 3.3% |
| District Courts | 569 | 4.5% |
| Total | 12,615 | 100.0% |

Our data set covers all Indian states and union territories, with rulings concentrated in industrialized regions and major metropolitan areas. The rulings cover the full spectrum of environmental issues, from industrial pollution and waste management to forest conservation and wildlife protection. Table 1 presents the most frequently cited environmental statutes in our dataset, providing insight into the main areas of environmental litigation.

However, despite its size and geographic coverage, our data set has several limitations. First, coverage of lower court decisions may be incomplete, particularly for earlier periods. Second, we only include rulings explicitly citing identified environmental statutes, potentially missing those that rely on other legal provisions. Third, the quality of case documentation has improved with time due to better digital practices. Finally, our data set reflects only rulings that reach formal adjudication, excluding disputes resolved through alternative mechanisms.

Despite these limitations, our dataset represents the most comprehensive compilation of Indian environmental court rulings available for research, providing unprecedented scope for analyzing judicial approaches to environmental protection over decades.

3. Methods

Our methodology involves four distinct phases: constructing the complete data set, establishing human-coding benchmarks, implementing the Large Language Model (LLM) analysis, and then analyzing model performance.

3.1. Dataset Construction

Our data set contains 12,615 environmental court rulings spanning 1974-2024, representing the full universe of litigation citing our identified statutes in the courts for which IndianKanoon has data. For computational efficiency and validation, we selected a subset of 1,910 rulings directly citing the Air

(Prevention and Control of Pollution) Act 1981, chosen because air pollution rulings represent a significant category of environmental litigation and this act is particularly salient in Indian environmental jurisprudence [3, 5].

3.2. Human Coding

During summer 2021, we recruited 14 law students from the National Law School of India in Bangalore to manually analyze the 1,910-case subset. All coders underwent comprehensive training through a detailed video guide and codebook. A senior research assistant supervised the entire process of allocating rulings to students, collecting responses, and monitoring the quality of the coding.

Each case was assigned to at least one coder, with 746 rulings (39%) receiving independent review by two coders to assess inter-rater reliability. When coders disagreed on the primary classification (pro-environment vs. not pro-environment), a third coder reviewed the case to determine the final classification. This occurred in only three rulings, indicating high inter-rater agreement.

The central question posed to human coders was: "Is this judgment likely to have a positive impact on the environment (or not)?" To answer the question, we provided additional guidance in the training manual.³ Reflecting a conservative approach that prioritizes direct, observable environmental interventions over potential indirect effects, this guidance directed coders to classify dismissed cases as having "no environmental impact."

3.3. LLM Models

Next we deployed two state-of-the-art Large Language Models (LLM) for further analysis: GPT-4 (via OpenAI API) and Claude 3.5 Sonnet (via Anthropic API). Implementation involved two distinct prompts, reflecting an evolution in our methodological approach.

Phase 1: Replication prompt Initially, we attempted to replicate the human coding process using the identical prompt given to human coders: *Is this judgment likely to have a positive impact on the environment (or not)?*"

Phase 2: Improved prompt After analyzing preliminary results and recognizing limitations in the original prompt, we developed an improved and more specific prompt: *Extract the result of the order. Respond 1 if the case likely has a near-term or immediate positive environmental impact that would reduce air pollution, otherwise respond 0 and do not write anything else.*

We modified the prompt for several methodological reasons. The first rationale was specificity. The improved prompt focuses on "near-term or immediate" impacts rather than general environmental effects, providing clearer evaluation criteria. The second was measurability. By specifying "reduce air pollution," the prompt targets a concrete, observable outcome rather than an abstract environmental benefit. Our third rationale was bias reduction. The revised prompt eliminates explicit instructions on dismissed cases, allowing the LLM to make more nuanced interpretations of the outcomes of the case. Finally, we note that the improved prompt provides more objective criteria, reducing subjective interpretation variability.

Both GPT-4 and Claude 3.5 Sonnet processed the same 1,910 rulings using both prompt versions with systematic quality controls including standardized API calls, error handling, and response validation. Due to minor technical issues (API timeouts, formatting errors), our final analytical sample contains 1,906 rulings with complete data from all three coding approaches, representing 99.8% of our intended sample.

³The training manual instructed coders: "If you think that the judgment is likely to have a positive impact, select 'Yes' from the drop-down menu. For example, if the court orders that a polluting factory be shut down or imposes fines on the polluter, such a judgment is likely to have a positive impact on the environment. If, on the other hand, you believe that the judgment will have no impact or a negative impact on the environment, select 'No' from the dropdown menu. This may include judgments where the petition is dismissed without passing any further orders. Judgments, where the case is sent back to a lower court to be heard afresh without passing any orders on the merits of the case, will also fall into this category."

We evaluated the models using standard classification metrics (accuracy, precision, recall, F1 score, and Krippendorff’s alpha) in multiple dimensions. We also performed robustness analysis on multiple subsamples.

4. Results

Tables 3 and 4 present the main summary statistics of our analysis. We note that while humans code about 25% of orders as green, GPT-4 codes about one-third (35.4%) of orders as green (Table 3). GPT-4 achieves its highest agreement with human coders using human prompt, with 73.9% overall accuracy, 48.8% precision, and 68.5% recall (Table 4). Claude 3.5 Sonnet shows the opposite pattern, achieving better agreement with the improved prompt (71.4% accuracy, 0.620 F1 score) compared to the human prompt (62.9% accuracy, 0.509 F1 score).

Table 3
Summary Statistics

| <i>Human Coded Sample</i> | N | Mean | SD | Min | Max |
|---|--------|-------|-------|-----|-----|
| Green Verdict (Human coding) | 1,910 | 0.252 | 0.434 | 0.0 | 1.0 |
| Green Verdict (GPT4 coding – human prompt) | 1,906 | 0.354 | 0.478 | 0.0 | 1.0 |
| Green Verdict (GPT4 coding – improved prompt) | 1,904 | 0.486 | 0.500 | 0.0 | 1.0 |
| Green Verdict (Claude coding – human prompt) | 1,896 | 0.431 | 0.495 | 0.0 | 1.0 |
| Green Verdict (Claude coding – improved prompt) | 1,894 | 0.429 | 0.495 | 0.0 | 1.0 |
| Full Sample Green Verdict (GPT4 coding) | 12,607 | 0.350 | 0.478 | 0.0 | 2.0 |

When we compare the two prompts across LLM models, we see that GPT-4 classified fewer rulings as environmentally favorable when using the human prompt (35.4%) compared to the improved prompt (48.6%). However, Claude shows minimal sensitivity to prompt variation, classifying a similar proportion of rulings as green with both the improved prompt (42.9%) and the human prompt (43.1%). In the case of the GPT-4 model, this pattern initially appears counterintuitive, given that the improved prompt specifically asked whether a ruling would have “near-term or immediate positive environmental impact that would reduce air pollution” - a more restrictive criterion than the broader question of whether it would “have a positive impact on the environment (or not)”.

Upon examining the rulings driving this discrepancy, we found that the difference is explained by procedural rulings with ambiguous outcomes. For example, in rulings involving multiple polluter defendants, where only some parties were ordered to implement abatement measures while others were exempted, determining the overall environmental impact proved challenging under either prompt formulation. We note that such procedural rulings are more likely to be interim court orders than final judgments, accounting for 31% of our sample. Our results remain robust when excluding this entire category of rulings.

In general, these findings highlight that prompt engineering effects vary significantly between different LLM architectures, suggesting that optimal prompting strategies may need to be model-specific rather than universally applicable.

Table 5 presents detailed confusion matrices showing classification agreement and disagreement patterns between human coders and each LLM model. As noted earlier, both LLM models systematically identify more rulings as environmentally favorable compared to human coders. GPT-4 with the improved prompt shows 541 false positives (rulings humans coded as “not green” but GPT-4 coded as “green”) versus only 95 false negatives. This pattern persists in both models and prompts, suggesting fundamental differences in how AI systems and human experts evaluate the environmental impact.

Analysis of 25 randomly selected disagreement rulings reveals that LLMs and humans use fundamentally different evaluation frameworks. In all examined rulings, humans classified rulings as “not green” while LLMs classified them as “green.” Human coders appear to interpret the rulings pessimistically based on their experience with India’s environmental policy implementation challenges, while LLMs

Table 4
Accuracy Metrics for LLM Models vs. Human Coding

| Metric | GPT-4 | | Claude 3.5 Sonnet | |
|----------------------|-----------------|--------------|-------------------|--------------|
| | Improved Prompt | Human Prompt | Improved Prompt | Human Prompt |
| Precision | 0.415 | 0.488 | 0.548 | 0.448 |
| Recall | 0.802 | 0.685 | 0.713 | 0.590 |
| F1 Score | 0.547 | 0.570 | 0.620 | 0.509 |
| Overall Accuracy | 0.666 | 0.739 | 0.714 | 0.629 |
| Krippendorff's Alpha | 0.282 | 0.383 | 0.392 | 0.210 |

Table 5
Confusion Matrices: LLM vs. Human Coding

| (a) GPT-4 - Improved Prompt | | | | | (b) GPT-4 - Human Prompt | | | | |
|------------------------------|-----------------------|-------|-------|--|---------------------------|-----------------------|-------|-------|--|
| Human | GPT-4 Classification | | | | Human | GPT-4 Classification | | | |
| | Not Green | Green | Total | | | Not Green | Green | Total | |
| Not Green | 884 | 541 | 1,425 | | Not Green | 1,081 | 345 | 1,426 | |
| Green | 95 | 384 | 479 | | Green | 151 | 329 | 480 | |
| Total | 979 | 925 | 1,904 | | Total | 1,232 | 674 | 1,906 | |
| (c) Claude - Improved Prompt | | | | | (d) Claude - Human Prompt | | | | |
| Human | Claude Classification | | | | Human | Claude Classification | | | |
| | Not Green | Green | Total | | | Not Green | Green | Total | |
| Not Green | 905 | 362 | 1,267 | | Not Green | 825 | 451 | 1,276 | |
| Green | 176 | 439 | 615 | | Green | 254 | 366 | 620 | |
| Total | 1,081 | 801 | 1,882 | | Total | 1,079 | 817 | 1,896 | |

Notes: Rows represent human classifications, columns represent LLM classifications. Diagonal elements show agreement, off-diagonal elements show disagreement.

displayed systematic optimism about formal legal outcomes, perhaps due to a lack of contextual understanding of enforcement realities. For example, when a court prevented illegal threshing machine use (Kanoon ID 20982084), human coders anticipated continued unauthorized use despite the ruling, while GPT-4 focused on the formal legal barrier established by the court decision.

Table 6 examines the performance of the model in various sub-samples. All analyses use the human prompt for consistency. Here we see that GPT-4 consistently outperforms Claude in all sub-samples, with accuracy ranging from 70.43% to 83.23%. Both models perform best in rulings that do not involve a Pollution Control Board (PCB) action, suggesting that procedural enforcement rulings present particular challenges for AI interpretation. We also note that the LLM models perform less well in Supreme Court and NGT rulings, possibly because these pertain to more complex cases.

When we applied GPT-4 to the complete data set of 12,615 rulings (using the improved prompt), it classified 35.0% of the rulings as environmentally favorable. This estimate aligns closely with the 35.4% rate in our validation subset, suggesting consistency in AI classification patterns throughout the entire data set.

5. Discussion

Our analysis reveals both promising opportunities and important limitations for AI-assisted environmental law analysis [16]. LLM models achieved approximately 74% accuracy compared to human expert coding, demonstrating substantial potential to scale legal analysis. This performance is consistent with

Table 6
Accuracy Across Different Subsamples

| | N | LLM vs. Human Coding | | Inter-Model |
|---------------------------|------|----------------------|-----------------|-----------------------------|
| | | GPT-4 Accuracy | Claude Accuracy | Agreement (GPT-4 vs Claude) |
| Rulings after 1990 | 1698 | 75.18% | 62.82% | 68.72% |
| Complete case information | 1674 | 75.21% | 62.78% | 68.50% |
| Rulings > 300 words | 1606 | 74.67% | 61.84% | 67.37% |
| Air pollution focus | 1416 | 72.44% | 62.08% | 69.44% |
| Supreme Court & NGT | 206 | 70.43% | 59.83% | 73.80% |
| Delhi NCR region | 475 | 71.56% | 63.57% | 67.47% |
| No PCB action | 888 | 83.23% | 66.16% | 71.39% |

Notes: "Complete case information" includes rulings where participants were successfully identified. "Rulings > 300 words" excludes brief procedural orders. "No PCB action" includes rulings not involving Pollution Control Board enforcement actions.

previous computational legal studies [10], suggesting that such accuracy levels represent significant success in AI applications to complex legal tasks.

The most notable pattern in our results is that LLMs consistently identified more rulings as environmentally favorable compared to human experts. GPT-4 classified 35.4% of the rulings as "pro-environment" versus 25.2% by human coders. However, our findings reveal systematic differences in the way AI and human experts assess environmental impact. Although LLMs excel at identifying formal legal outcomes, human experts incorporate a contextual understanding of enforcement challenges that LLMs lack. The systematic patterns of disagreement, rather than random errors, suggest that these groups access fundamentally different types of information. This reveals that human judgment remains essential for evaluating implementation prospects, particularly in environmental law, where the gaps between judicial declarations and enforcement significantly affect real-world outcomes.

These results have several implications for legal research and policy analysis. AI tools offer unprecedented efficiency for systematic analysis of large legal datasets, enabling researchers to identify patterns and track judicial trends at previously impossible scales. For practitioners, AI could streamline legal research by helping to identify relevant precedents and litigation strategies. However, effective AI-assisted legal analysis requires acknowledging these limitations and developing hybrid approaches that combine computational efficiency with human expertise.

These insights inform future studies that seek to improve methodology and policy. Our approach demonstrates how human expert validation can be systematically integrated into AI-assisted legal research. Our documented disagreement patterns could guide the development of more sophisticated legal AI systems that incorporate enforcement probability models alongside formal legal analysis.

Finally, this research opens a new path for the analysis of the broad impacts of environmental policy in India. Our data set of 12,165 rulings provides a foundation for examining legal arguments and connecting judicial decisions to measurable environmental outcomes, as has been extensively conducted elsewhere ([18]). In future work, we hope to deploy LLM models to map rulings to specific geographic jurisdictions, extract key arguments from the corpus, and identify evolving trends in Indian environmental jurisprudence. By integrating court ruling data with pollution indicators, we aim to quantify the how judicial decisions impact environmental outcomes. Such research is particularly vital in the context of India, where pollution levels are quite severe, yet this type of large-scale analysis has not yet been conducted on a large scale [4, 6, 19].

6. Conclusion

This study demonstrates the potential of AI to improve the analysis of environmental court rulings, achieving 73% agreement with human coders on our comprehensive dataset of 12,615 Indian environmental rulings. Although AI effectively catalogs formal legal interventions and tracks doctrinal

developments, human judgment remains essential for evaluating implementation prospects and policy effectiveness. These findings suggest that hybrid approaches combining computational efficiency with human expertise can significantly improve the scalability of legal research, particularly where administrative data are not standardized, opening new avenues for revolutionizing the analysis of large-scale legal datasets across jurisdictions and policy domains.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 and Claude Sonnet 4 for grammar and spelling checks, as well as paraphrasing and rewording assistance. After using these services, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

- [1] UNEP: Environmental Courts and Tribunals – 2021: A Guide for Policy Makers (2022). <https://wedocs.unep.org/20.500.11822/40309>
- [2] Rajamani, L.: Public Interest Environmental Litigation in India: Exploring Issues of Access, Participation, Equity, Effectiveness and Sustainability. *Journal of Environmental Law* 19(3), 293–321 (2007).
- [3] Bhuvania, A.: *Courting the people: Public interest litigation in post-emergency India*, vol. 2. Cambridge University Press (2017).
- [4] Gill, G.N.: *Environmental Justice in India: The National Green Tribunal*. Routledge, London (2017).
- [5] Ghosh, S. (ed.): *Indian Environmental Law: Key Concepts and Principles*. Orient BlackSwan (2019).
- [6] Do, Q.-T., Joshi, S., Stolper, S.: Can environmental policy reduce infant mortality? Evidence from the Ganga Pollution Cases. *Journal of Development Economics* 133, 306–325 (2018)
- [7] Chandra, A., Kalantry, S., Hubbard, W.H.J.: *Court on Trial: A Data-Driven Account of the Supreme Court of India*. Penguin Random House India (2023)
- [8] Ash, E., Asher, S., Bhowmick, A., Bhupatiraju, S., Chen, D., Devi, T., Goessmann, C., Novosad, P., Siddiqi, B.: In-group bias in the Indian judiciary: Evidence from 5 million criminal cases. *Review of Economics and Statistics*, 1–45 (2025)
- [9] Bhupatiraju, S., Chen, D.L., Joshi, S.: The Promise of Machine Learning for the Courts of India. *Nat'l L. Sch. India Rev.* 33, 463 (2021).
- [10] Katz, D.M., Bommarito II, M.J., Blackman, J.: A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS One* 12(4), e0174698 (2017)
- [11] Athey, S., Imbens, G.W.: Machine learning methods that economists should know about. *Annual Review of Economics* 11, 685–725 (2019).
- [12] Horton, J.J.: Large language models as simulated economic agents: What can we learn from homo silicus? *National Bureau of Economic Research* (2023).
- [13] Kim, J., Lee, J., Jang, K.M., Lourentzou, I.: Exploring the limitations in how ChatGPT introduces environmental justice issues in the United States: A case study of 3,108 counties. *Telematics and Informatics* 86, 102085 (2024).
- [14] Korinek, A.: Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature* 61(4), 1281–1317 (2023).
- [15] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large language models transform computational social science? *Computational Linguistics* 50(1), 237–291 (2024).
- [16] Re, R.M., Solow-Niederman, A.: Developing artificially intelligent justice. *Stanford Technology Law Review* 22, 242 (2019).
- [17] Bhupatiraju, S., Chen, D.L., Joshi, S., Neis, P.: Impact of free legal search on rule of law: Evidence from Indian Kanoon (2024). Unpublished manuscript.
- [18] Liepiņa, R., Meyer-Erdmann, M., Serrano, P. H., Simoncini, W. (2024). Computational annotation

and database developments for European works Councils and law. In *Field Guide to Researching Employment and Industrial Relations* (pp. 141-160). Edward Elgar Publishing.

- [19] Bhupatiraju, S., Chen, D.L., Joshi, S., Neis, P. and Singh, S.: *Environmental Litigation as Scrutiny: A Four Decade Analysis of Justice, Firms, and Pollution in India* (October 31, 2024). Available at SSRN. Paper 5013333.