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

Patrick Behrer[1][0000−0001−6946−7213], Shareen Joshi[2][0000−0001−5693−7140], Olexiy Kyrychenko[3][0000−0002−3557−7892], Viknesh Nagarathinam[4][0009−0003−3126−5589], Peter Neis[5][0009−0008−4358−4110], and Shashank Singh[6][0009−0003−0692−3850]

[1] World Bank, Washington DC, USA abehrer@worldbank.org
[2] Georgetown University, Washington DC, USA sj244@georgetown.edu
[3] Radboud University, the Netherlands olexiy.kyrychenko@ru.nl
[4] Georgetown University, Washington DC, USA viknesh.n91@gmail.com
[5] Université Clermont Auvergne, CNRS, IRD, CERDI, France peter.neis@uca.fr
[6] The University of Chicago, Illinois, USA shashank.singh@chicagobooth.edu

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 cases 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 cases. 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 · 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 policymakers. This issue is particularly acute in India, where the judiciary has emerged as a global leader in environmental governance [2–4] but empirical analysis of decisions has been quite limited [7].

The analysis of environmental rulings faces some fundamental limitations: Manual review of thousands of unstructured legal documents is cumbersome and requires specialized expertise [10]. Coding large number of cases systematically and consistently can be prohibitively expensive. Recent advances in Large Language Models (LLMs) offer a promising solution to these issues, demonstrating strong capabilities in the analysis of complex legal texts [12–17].

This study examines whether AI systems perform as well as human experts in assessing whether judicial decisions produce positive environmental outcomes in the context of India. This determination is crucial in deciding whether we can expand the analysis of environmental jurisprudence to inform policy and improve access to environmental justice.

We examine a novel data set of 12,615 environmental court cases 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 cases. Our central task, determining whether a judicial decision is "pro-environment," captures 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[12]. Second, we create the first comprehensive AI-annotated dataset of 12,615 Indian environmental cases, 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. We conducted a comprehensive search of the Indian Kanoon.org database, identifying 2,996 judicial rulings that explicitly cited at least one of these acts[7] 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.

Our final dataset encompasses all judicial rulings from 1974 onward citing any identified environmental statute, resulting in 12,615 court cases spanning through 2024. Most cases originated in the High Courts (69%), followed by the National Green Tribunal (23.1%) and the Supreme Court (3.3%). Document

[7] IndianKanoon.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 [11].

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.

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

As noted above, our data set contains 12,615 environmental court cases that spanned the years 1974-2024. This comprehensive data set serves as the foundation for our analysis and represents the full universe of environmental litigation citing our identified statutes. For computational efficiency and validation purposes, we selected a subset of 1,910 cases that directly cited the Air (Prevention and Control of Pollution) Act 1981. This subset was chosen because air pollution cases represent a significant and well-defined category of environmental litigation. Moreover, the Air Act is one of India's foundational environmental statutes.

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, allocating cases, and monitoring coding quality.

Each case was assigned to at least one coder, with 746 cases (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 3 cases, 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.[8] This approach instructed

[8] The 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 drop-down 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."

human coders to classify dismissed cases as having "no environmental impact," which we later recognized as potentially introducing bias, as some dismissals might indirectly result in environmental benefits.

3.3 LLM Models

We deployed two state-of-the-art Large Language Models for our 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, 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 cases 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 cases with complete data from all three coding approaches, representing 99.8% of our intended sample.

We evaluated the models using standard classification metrics (accuracy, precision, recall, F1 score, and Krippendorff's alpha) across multiple dimensions. We also performed robustness analyzes on multiple sub-samples.

4 Results

Tables 1 and 2 present the main summary statistics of our analysis. We note that GPT-4 achieves its highest agreement with human coders using human prompt, with 73.9% overall accuracy, 48.8% precision, and 68.5% recall. 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 1: Summary Statistics

Human Coded Sample	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 cases 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 cases 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 cases driving this discrepancy, we found that the difference is explained by procedural rulings with ambiguous outcomes. For example, in cases 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 noted 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 3 presents detailed confusion matrices showing classification agreement and disagreement patterns between human coders and each LLM model. We note that both LLM models systematically identify more cases as environmentally favorable compared to human coders. GPT-4 with the improved prompt shows 541 false positives (cases 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.

Table 2: Accuracy Metrics for LLM Models vs. Human Coding

Metric	GPT-4 Improved Prompt	GPT-4 Human Prompt	Claude 3.5 Sonnet Improved Prompt	Claude 3.5 Sonnet 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 3: Confusion Matrices: LLM vs. Human Coding

(a) GPT-4 - Improved Prompt

Human	GPT-4 Classification Not Green	GPT-4 Classification Green	Total
Not Green	884	541	1,425
Green	95	384	479
Total	979	925	1,904

(b) GPT-4 - Human Prompt

Human	GPT-4 Classification Not Green	GPT-4 Classification Green	Total
Not Green	1,081	345	1,426
Green	151	329	480
Total	1,232	674	1,906

(c) Claude - Improved Prompt

Human	Claude Classification Not Green	Claude Classification Green	Total
Not Green	905	362	1,267
Green	176	439	615
Total	1,081	801	1,882

(d) Claude - Human Prompt

Human	Claude Classification Not Green	Claude Classification Green	Total
Not Green	825	451	1,276
Green	254	366	620
Total	1,079	817	1,896

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

Analysis of 25 randomly selected disagreement cases reveals that LLMs and humans use fundamentally different evaluation frameworks. In all examined cases, humans classified rulings as "not green" while LLMs classified them as "green." Human coders interpreted the rulings pessimistically based on their experience with India's environmental policy implementation challenges, while LLMs displayed systematic optimism about formal legal outcomes, clearly lacking a contextual understanding of enforcement realities. For example, when a court prevented illegal thresher 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 4 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 cases that do not involve a Pollution Control Board (PCB) action, suggesting that procedural enforcement cases present particular challenges for AI interpretation. We also note that the LLM models perform less well in Supreme Court and NGT cases, possibly because these are more complex cases.

Table 4: Accuracy Across Different Subsamples

	N	LLM vs. Human Coding		Inter-Model
		GPT-4 Accuracy	Claude Accuracy	Agreement (GPT-4 vs Claude)
Cases after 1990	1698	75.18%	62.82%	68.72%
Complete case information	1674	75.21%	62.78%	68.50%
Cases > 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 cases where participants were successfully identified. "Cases > 300 words" excludes brief procedural orders. "No PCB action" includes cases not involving Pollution Control Board enforcement actions.

When we applied GPT-4 to the complete data set of 12,615 cases (using the improved prompt), it classified 35.0% of the cases 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 [18]. LLM models achieved approximately 74% accuracy compared to human expert coding, demonstrating substantial potential to scale legal analysis. This performance is consistent with previous computational legal studies [12], suggesting that such accuracy levels represent significant success in AI applications to complex legal tasks.

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 gaps between judicial declarations and enforcement significantly affect real-world outcomes.

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, our comprehensive dataset of 12,615 environmental cases provides a valuable resource for future research on environmental jurisprudence in India. This type of research is particularly crucial in India's context, where severe air and water pollution challenges make effective environmental governance essential for public health and sustainable development [2–4, 6, 7].

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 cases. 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.

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. Bhuwania, 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. Baxi, U.: The Aleatory of Justice: Poverty, Democracy and Human Rights. Oxford University Press (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. Ghosh, S. (ed.): Indian Environmental Law: Key Concepts and Principles. Orient BlackSwan (2019).
8. 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)
9. 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)
10. 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).
11. Bhupatiraju, S., Chen, D.L., Joshi, S., Neis, P.: Impact of free legal search on rule of law: Evidence from Indian Kanoon (2024).
12. 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)
13. Athey, S., Imbens, G.W.: Machine learning methods that economists should know about. Annual Review of Economics 11, 685–725 (2019).
14. Horton, J.J.: Large language models as simulated economic agents: What can we learn from homo silicus? National Bureau of Economic Research (2023).
15. 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).
16. Korinek, A.: Generative AI for economic research: Use cases and implications for economists. Journal of Economic Literature 61(4), 1281–1317 (2023).
17. 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).
18. Re, R.M., Solow-Niederman, A.: Developing artificially intelligent justice. Stanford Technology Law Review 22, 242 (2019)