Decoding green justice: AI-assisted analysis of environmental court rulings in India

A. Patrick Behrer, Daniel L. Chen, Shareen Joshi, Olexiy Kyrychenko, Viknesh Nagarathinam, Peter Neis and Shashank Singh

- AI-assisted methods can analyse large-scale legal datasets with approximately 70% accuracy, enabling more effective monitoring of environmental litigation outcomes.
- Large language models (LLMs) like ChatGPT-4 can assess whether court rulings have a positive environmental impact with accuracy rates that approach human expert analysis.
- Analysis of 12,615 environmental court cases in India reveals that approximately 35% of rulings are intended to be favourable to the environment, with significant variations across courts and case types.

POLICY BRIEF IND-23019

FEBRUARY 2025

This project was funded by the IGC

theigc.org

DIRECTED BY

FUNDED BY

India faces severe environmental challenges, with 21 of the world's 30 most polluted cities within its borders (IQAir, 2023). Despite robust environmental legislation since the 1970s, implementation remains problematic, with the Air Quality Index regularly reaching "severe" levels in metropolitan areas (Greenstone and Fan, 2020). The judiciary has emerged as a proactive force in environmental governance through Public Interest Litigation (PIL) mechanisms, with the Supreme Court and the specialised National Green Tribunal (NGT; established in 2010) issuing landmark rulings. However, until now, researchers have struggled to systematically analyse the impact of these court interventions due to the challenge of harnessing legal data for empirical analysis (Bhupatiraju et al., 2021).

Harnessing AI for legal environmental analysis

The digitisation of judicial records has created new opportunities for evidence-based policymaking, but researchers face significant challenges in harnessing these large datasets. Traditional methods of legal analysis are limited by the complexity of legal data and the reality of inconsistent data formats across court systems. Data from the Indian judiciary, available through e-court systems and court websites, often lacks consistent tagging of case numbers, key dates, and actors (Bhupatiraju et al., 2021). Most empirical studies have typically been limited to analysing a small set of variables or focusing on small subsamples of cases (Do et al., 2018; Rao, 2018, 2021; Bhupatiraju et al., 2024).

Advanced AI algorithms, particularly large language models (LLMs), have shown considerable promise in other settings (Athey and Imbens, 2019; Horton, 2023, Korinek, 2023). This research demonstrates that LLMs can effectively summarise and code environmental court cases at scale.

Our study analysed 12,615 environmental court orders in India spanning three decades, using both human coders and AI models to assess case outcomes. We compare two state-of-the-art LLMs (OpenAI's GPT-4 and Anthropic's Claude 3.5 Sonnet) against a subset of 1,905 cases manually labelled by law students, providing a robust benchmark for assessing the capabilities of LLMs in specialised legal domains.

Key findings

We compare the performance of both LLM models to humans in the sample of 1,910 human-coded cases (Figure 1). Human analysis classified 25.2% of rulings as pro-environment ("green"). AI models showed a greater tendency to identify positive environmental outcomes - ChatGPT-4 initially classified 48.6%

of cases as green, dropping to 35% when using identical prompts to humans. Claude classified 42.9% of cases as green, slightly increasing to 43.1% with human-equivalent prompts. This consistent pattern suggests AI models are more inclined to interpret Indian environmental rulings favorably than human experts (See Box 1). We also explored the accuracy of LLM models in sub-samples:

- **ChatGPT-4 demonstrated robust performance** (as defined by predictions in line with human coders) with accuracy ranging from 75% to 84%, with the highest accuracy (83.23%) in cases without Pollution Control Board involvement.
- **The ChatGPT-4 model maintained strong performance across multiple dimensions**: cases from later years, those with clearly identified parties and judges, substantive cases exceeding 300 words, specialised air pollution cases, Supreme Court and National Green Tribunal (NGT) jurisdictions, and Delhi National Capital Region (NCR region) cases.
- **Claude was less likely to match human coders than ChatGPT-4 across all subsamples**. When comparing the two models, Claude's predictions aligned with ChatGPT-4 between 68-74% of the time, with the strongest agreement (89.30%) in cases heard at the Supreme Court or NGT.

FIGURE 1: Comparison of ChatGPT-4 with humans

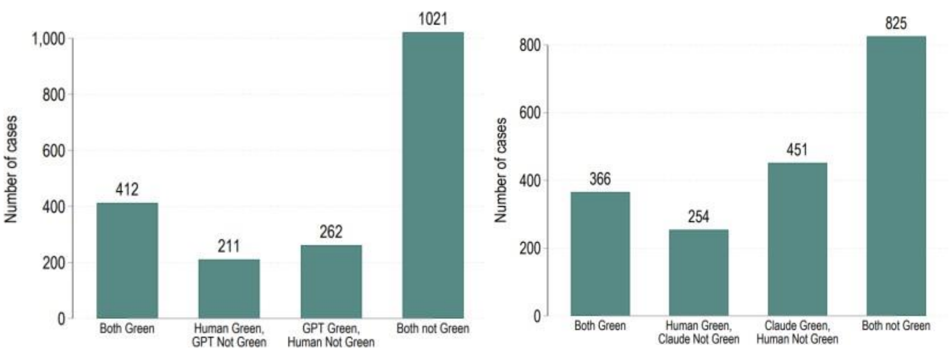

TABLE 1: Summary statistics, human sample

Human Coded Sample (N=1910)	Mean	SD
Green Verdict (Human coding)	0.252	0.434
Green Verdict (GPT4 – human prompt)	0.354	0.478
Green Verdict (GPT4 – improved prompt)	0.486	0.500
Green Verdict (Claude – human prompt)	0.431	0.495
Green Verdict (Claude – improved prompt)	0.429	0.495
Number of human readers	1.440	0.500
Sum of scores of human readers	0.371	0.561

Appeal case (Human coding)	0.283	0.450
Constitutional case (Human coding)	0.127	0.333
Govt plays a role (Human coding)	0.000	0.000
Case Relevant to the env (Scale 0-2)	0.830	0.376
PCB Action (GPT4)	0.472	0.499
Regulator Action (GPT4)	0.564	0.496
Length of case (characters)	4915	11705
Delhi NCR Region	0.282	0.450

TABLE 2: Summary statistics, expanded sample

Expanded sample (Coded by ChatGPT-4 (N=12,615))	Mean	SD
Green Verdict (GPT4 coding)	0.350	0.478
Green Verdict (human coding)	0.314	0.465
Order	0.199	0.400
Regulator Action (GPT4)	0.357	0.479
PCB Action (GPT4 coding)	0.273	0.446
Politician Action (GPT4 coding)	0.045	0.207
Number of petitioners	2.119	6.084
Number of respondents	3.102	6.256
Number of judges	1.534	0.918
Number of states	1.065	0.941
Supreme Court case (GPT4 coding)	0.032	0.177
High Court case (GPT4 coding)	0.689	0.463
NGT case (GPT4 coding)	0.226	0.418
Delhi NCR Region	0.290	0.454

TABLE 3: Additional statistics for accuracy

Panel (a): Common Cases, LLM models versus Human prompt	N	GPT4 Accuracy	N	Claude Accuracy
All cases in this sample	1906	75.18%	1896	62.82%
Cases after 1990	1906	75.18%	1896	62.82%
Cases with 1+ petitioner, judge and respondent	1880	75.21%	1870	62.78%
Cases that are greater then 300 words	1800	74.67%	1790	61.84%
Cases relevant to air pollution	1582	72.44%	1577	62.08%
Cases heard at the Supreme Court and Green Tribunal	230	70.43%	229	59.83%
Cases in the Delhi NCR Region	538	71.56%	538	63.57%
Cases featuring no action by the PCB	1002	83.23%	996	66.16%

Panel (b): Common cases, ChatGPT-4 compared to Claude		
All cases in this sample	1896	68.72%
Cases after 1990	1896	68.72%
Cases with 1+ petitioner, judge and respondent	1870	68.50%
Cases that are greater then 300 words	1790	67.37%
Cases relevant to air pollution	1577	69.44%
Cases heard at the Supreme Court and Green Tribunal	229	73.80%
Cases in the Delhi NCR Region	538	67.47%
Cases featuring no action by the PCB	996	71.39%

Policy implications

We are currently using our data to examine the link between environmental court cases and air pollution levels in Delhi and, eventually, all of India. We are combining our legal dataset with granular air quality measurements and meteorological controls to estimate the initial impact of judicial effectiveness in this context. Our dataset can also be leveraged by policymakers for improving environmental governance in India:

1. **Monitoring implementation gaps**: By tracking outcomes systematically, policymakers can identify where court orders are green and how these correlate with (or do not correlate with) environmental improvements.
2. **Judicial education**: Analysis reveals how different benches approach environmental evidence, potentially harmonising jurisprudence across India's complex judicial landscape.
3. **Accountability**: Greater transparency in environmental rulings enables civil society to hold authorities accountable for implementing court orders.
4. **Policy design**: Understanding patterns in judicial outcomes can inform more effective environmental regulation design.

This methodological approach has applications beyond India. As courts worldwide increasingly digitise their records, AI-assisted analysis could allow more empirical studies of the link between environmental jurisprudence and real world outcomes. For countries struggling with environmental issues or climate change, this approach offers a new lens to examine the judiciary's role in environmental stewardship.

While AI offers tremendous potential for scaling environmental justice monitoring, optimal results require combining AI efficiency with human understanding of context. AI excels at processing formal outcomes at scale, while humans bring crucial contextual knowledge about implementation realities. Together, they provide a more complete picture.

BOX 1: Human vs. machine in assessing environmental justice

Our analysis reveals a significant divergence between how AI models and human experts evaluate environmental court rulings. While achieving 70% overall agreement, AI consistently identified more environmentally favourable rulings than human coders (35-48% vs. 25% "green" rulings).

On the one hand, we can expect some human cynicism. Humans, familiar with India's implementation challenges, frequently rated seemingly positive rulings as ineffective, anticipating enforcement failures. AI, on the other hand, focuses on formal outcomes without considering practical limitations. For instance, in a case preventing the use of an illegal polluting machine, human coders classified it as having no environmental impact, anticipating continued unauthorised use despite the court's intervention. ChatGPT-4, focusing on formal outcomes, coded this as environmentally positive.

These findings suggest that while AI offers tremendous potential for scaling up environmental justice monitoring across vast legal datasets, optimal results require combining AI efficiency with human understanding of context

References

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*, 685–725.

Bhupatiraju, S., Chen, D. L., Joshi, S., & Neis, P. (2024). Impact of free legal search on rule of law: Evidence from Indian Kanoon. *In*.

Bhupatiraju, S., Chen, D. L., Joshi, S., Neis, P., & Singh, S. (2024). Litigation as Scrutiny: A Four Decade Analysis of Environmental Justice, Firms, and Pollution in India. *In*.

Bhupatiraju, S., Chen, D. L., & Joshi, S. (2021). The Promise of Machine Learning for the Courts of India. *National Law School of India Review, 33*, 463.

Central Pollution Control Board. (2021). Biological water quality assessment of the River Ganga (2020-21). Ministry of Environment, Forest and Climate Change, Government of India.

Damle, D., & Anand, T. (2020). Problems with the e-Courts data. *National Institute of Public Finance and Policy Working Paper, 314*.

Do, Q.-T., Joshi, S., & Stolper, S. (2018). Can environmental policy reduce infant mortality? Evidence from the Ganga Pollution Cases. *Journal of Development Economics, 133*, 306–325.

Greenstone, M., & Fan, C. Q. (2020). Air Quality Life Index Annual Update. Energy Policy Institute at the University of Chicago.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *National Bureau of Economic Research*.

IQAir. (2023). World Air Quality Report 2022: Region and City PM2.5 Ranking. IQAir AG.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature, 61*(4), 1281–1317.

Motoki, F., Neto, V. P., & Rodrigues, V. (2024). More human than human: measuring ChatGPT political bias. *Public Choice, 198*(1), 3–23.

Rao, M. (2021). Courts redux: Micro-evidence from India. *Working Paper*.

Rozado, D. (2023). The political biases of chatgpt. *Social Sciences, 12*(3), 148.