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

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

April 23, 2025

1 Introduction

The proliferation of digitized public data has revolutionized evidence-based policy making (Geiger and Lucke 2011; Einav and Levin 2014; Haskins 2018; Penner and Dodge 2019). Courts are no exception to this trend (Susskind 2019; Liang 2023). Judicial systems around the world have digitized records, established e-justice platforms, and streamlined information exchange within courts and with their users (Reiling and Contini 2022; Moraes, Lunardi, and Correia 2024).

However, researchers face significant challenges in harnessing the growing corpus of judicial information (Geiger and Lucke 2011). The available information is frequently fragmented, inconsistently formatted, and replete with domain-specific terminology, due to the complex organizational and professional protocols that underlie it (Posner 2014). Navigating this landscape without legal experience presents a significant challenge.

*Behrer: Development Economics Research Group, World Bank, 1818 H Street Washington, DC 20433 (e-mail: abehrer@worldbank.org); Chen: Toulouse School of Economics, Université Toulouse Capitole, Toulouse, France (e-mail: daniel.chen@iast.fr); Joshi: Walsh School of Foreign Service, Georgetown University, 3700 O Street, Washington DC, 20057 (e-mail: shareen.joshi@georgetown.edu); Kyrychenko: Nijmegen School of Management, Radboud University, Heyendaalseweg 141, 6525 AJ Nijmegen, the Netherlands (e-mail: olexiy.kyrychenko@ru.nl); Nagarathinam: [ADD DETAILS HERE]; Neis: Université Clermont Auvergne, CNRS, IRD, CERDI, F-63000, Clermont-Ferrand, France (e-mail: peter.neis@uca.fr); Singh: The University of Chicago Booth School of Business (e-mail: shashank.singh@chicagobooth.edu). 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.

Recent studies have highlighted the potential for Artificial Intelligence (AI) and Machine Learning (ML) techniques to enhance the summarization and interpretability of large-scale administrative data (Athey and Imbens 2019; Aher, Arriaga, and Kalai 2023; Horton 2023; Korinek 2023; Ziems et al. 2024). We explore these methods in the context of India, where a large amount of administrative data has become publicly available over the past decades (Bhupatiraju, Chen, and Joshi 2021; Bhupatiraju, Chen, Joshi, and Neis 2024; Ash et al. 2022). The Indian judiciary presents unique challenges, including lengthy judgments, inconsistent data quality, and lack of tractability (Damle and Anand 2020).[1] Consequently, empirical studies have typically been limited to analyzing a small set of variables or simply focus on small subsamples (Chandra, Hubbard, and Kalantry 2017; Do, Joshi, and Stolper 2018; Rao 2021; Bhupatiraju, Chen, Joshi, Neis, and Singh 2024).

We apply advanced AI algorithms to curate and code a comprehensive dataset of 12,615 environmental court cases in India. Using state-of-the-art large language models (LLMs), specifically generative pre-trained transformers (GPT), we summarize court rulings to assess their environmental impact.[2] Our primary tool is OpenAI's GPT-4 model, widely known as ChatGPT-4. To evaluate GPT-4 performance, we use a manually labeled subset of 1,905 cases, analyzed by human annotators. Furthermore, we compare ChatGPT-4 with Anthropic's Claude 3.5 Sonnet model. Finally, we assess the robustness and accuracy of GPT-4 across various data sub-samples relevant to economists and policymakers.

This work offers two main contributions. First, we present a novel dataset of environmental cases adjudicated by the Indian judiciary, created through a blend of human coding and LLM models. Given India's status as one of the most polluted countries in the world and the proactive role of its judiciary in shaping environmental policy, this data set provides valuable new opportunities for economists, data scientists, journalists and policy makers interested in environmental governance and legal impacts.

Our second contribution advances the field of AI in law. We demonstrate the high accuracy achievable by GPT-4 in a specialized legal domain on a large data set. Specifically, both ChatGPT-4 and Claude achieve around 70% accuracy relative to human annotators, with ChatGPT-4 outperforming Claude. These results align with previous findings on smaller datasets Savelka et al. 2023 and highlight the potential of AI models for legal analysis in specific areas.

The paper is organized as follows. Section 2 details our data, Section 3 outlines our research methods, Section 4 provides the results, including some robustness checks. Section 5 discusses our key findings. The final section concludes.

[1] Data from the Indian judiciary, available through e-courts systems and court websites, often lacks consistent tagging of case numbers, key dates, actors, and case status.

[2] Large language models (LLMs) are a type of artificial intelligence (AI) model that utilizes machine learning (ML) to process and generate human-like text (Brown et al. 2020).

2 Data

Our research began with a comprehensive review of Indian environmental laws.[3] We identified three significant acts: (1) the Water (Prevention and Control of Pollution) Act 1974 (Water Act), (2) the Air (Prevention and Control of Pollution) Act 1981 (Air Act), and (3) the Environment (Protection) Act 1986 (EP Act). From a full scrape of the free public database IndianKanoon.org, we found 2,996 judicial rulings that had cited these three acts. A careful examination of additional laws cited in this corpus of cases revealed 23 additional environmental regulations that were cited in these rulings.[4] We further expanded our dataset to encompass all judicial rulings that referenced any of these additional legislative acts. A list of acts that were cited by at least 200 cases is presented in Table 1.

Our final sample consisted of 12,615 rulings. Of these, 8,706 (69%) came from the High Courts, 2,925 (23.1%) came from the Green Tribunal of India, which was established in 2010, 415 (3.29%) came from the Supreme Court of India, and the remainder came from district courts.

The primary obstacle in analyzing this sample of cases, which are a mix of final judgments and interim orders, is its unstructured nature. The corpus of 12,615 rulings consisted of diverse, unformatted text documents, each with a unique form of identification from the website where we scraped the cases. Our priority in coding the cases was thus to extract some basic attributes from these files. These included an official case number, the names of litigants (petitioners, respondents and judges), the type of case underlying the ruling (criminal or civil), characteristics of the case (judgment or order), and locations where the final ruling was relevant. The broad objective of this structured approach was the creation of a database for the systematic analysis of the environmental cases.

3 Methods

Our research proceeded in several steps. We first manually read and coded 1,910 legal cases that directly cited the Air (Prevention and Control of Pollution) Act 1981 (Air Act).

[3]For this we relied on desk research and also conversations with Shibani Ghosh, Mrinal Satish and two prominent environmental lawyers in India, whom we spoke to in 2021. We drew heavily from Ghosh (2019) for a history of environmental legislation in India.

[4]We retained all the additional citations in these rulings, even if they had just been cited once. The full list includes the following: E-Waste (Management) Rules 2016, Batteries (Management & Handling) Rules 2001, Battery Waste Management Rules 2020, Bio-Medical Waste Management Rules 2016, Plastic Waste Management Rules 2016, Solid Waste Management Rules 2016, Construction and Demolition Waste Management Rules 2016, Hazardous and Other Waste (Management and Transboundary Movement) Rules 2016, Manufacture, Storage and Import of Hazardous Chemicals Rules 1989 (MSIHC Rules), Coastal Regulation Zone Notification 2019 (and related 2021 procedure for violation of the CRZ Notification), Environment Impact Assessment Notification 2006, Wild Life (Protection) Act 1972, Forest (Conservation) Act 1980, Public Liability Insurance Act 1991, Biological Diversity Act 2002, National Green Tribunal Act 2010, Section 91 of the Civil Procedure Code, The Water (Prevention and Control of Pollution) Cess Act, 1977, The Forest (Conservation) Act, 1980, The Air (Prevention and Control of Pollution) Act, 1981, The Environment (Protection) Act, 1986, The Public Liability Insurance Act, 1991, The Biological Diversity Act, 2002.

Next, we used ChatGPT-4 and Claude to read these cases as well as the remaining cases in the sample, likewise coding their attributes and assessing their potential environmental impact. We describe each of these coding efforts below.

3.1 Human Coding

During the summer of 2021, we recruited a team of 14 law students at the National Law School of India in Bangalore, India to analyze the 1,910 air pollution cases. Under the supervision of a senior research assistant, who both allocated cases and monitored quality, the students underwent training via a comprehensive video guide based on a detailed Codebook (available in the Appendix). Following their training, each student was assigned 10-15 cases for independent review. They conducted their analysis and documented their findings by completing an online form for each case.

Over a period of three months, all 1,910 cases were read by one person. 746 (39%) cases were read by two people. When there was a difference in opinion on whether a case was pro-environment, a third student would read the case to break the tie.[5]

The most significant question asked of the students was whether the judgment was pro-environment (or rather, green). The specific prompt we provided was as follows:

> *Is this judgment likely to have a positive impact on the environment (or not)?*

In the training manual, we provided some further guidance on how they should answer this question. We wrote the following:

> *In this field, we want to determine whether the judgment is likely to have a positive impact on the environment or not. 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 for being heard afresh without passing any orders on the merits of the case, will also fall in this category.*

3.2 ML Coding

In the summer of 2024, we deployed a large language model, ChatGPT-4, to answer several questions about the full set of environmental rulings (included in the Appendix). We ran

[5] A tie occurred in only 3 cases.

this algorithm on the full set of 12,615 cases. For this expanded sample, we improved the prompt to assess whether a verdict was pro-environment by adopting the following phrasing:

> *Extract the outcome 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 believe that by focusing on a specific aspect of case impact – reducing air pollution in the near-term – this prompt provides a clearer, more objective criterion for evaluation, reducing subjectivity and variability in responses. We also eliminated the language on handling dismissals, allowing the algorithm to draw deeper inferences and circumvent some of the biases that human coders had shown.[6]

We also designed additional prompts to capture the attributes of the case. We included an additional question asking if the ruling was an order or a final judgment.[7] We expanded the question of the role of the government in the case to specifically ask whether the case featured action by pollution control boards (regulators) or politicians.

We also included additional questions about the location of the jurisdiction of the case, with the added caveat that the location of the court may be the same as that of the jurisdiction. The full list of questions is presented in the appendix of this paper.

3.3 Additional ML Model: Claude 3.5 Sonnet

As an additional robustness check, we ran the same set of prompts that were used on ChatGPT-4 through Claude 3.5 Sonnet, another large language model developed by Anthropic. Claude processed the same 1,905 cases using identical prompts as those used with ChatGPT-4 as well as humans. This parallel analysis allowed us to assess the consistency of our results across different LLM architectures and training approaches and compare these to humans.

4 Results

We present summary statistics of human and LLM coded samples in Table 2. In the sample of 1,910 human-coded cases, 25.2% of the rulings were classified as pro-environment

[6] In the original coding manual (see Appendix), human coders were instructed to code dismissed cases as having "no environmental impact". Upon reflection, we recognize this as a potential source of bias, as some dismissals (e.g., petitions by firms against regulatory inspectors) may have indirectly resulted in environmental action. This revised approach allows for a more nuanced interpretation of dismissed cases.

[7] Orders are typically brief directives issued during proceedings or to address specific procedural matters; they can be interim and subject to modification. In contrast, judgments are comprehensive final decisions that conclude a case, providing detailed reasoning and determining the substantive rights of the parties. Orders can be issued at various stages of a legal proceeding, while judgments are usually delivered at the conclusion. Judgments can generally be appealed in higher courts, whereas orders can rarely be an appeal. The format and content of judgments are thus typically more formal relative to orders.

(henceforth, we refer to this as "green"). When examining the subset of cases that underwent multiple human reviews (not shown here), this percentage decreased slightly to 19. 8%.

In this limited sample of common cases, we find that ChatGPT-4 and Claude classified 48.6% and 42.9% of the rulings as green (Table 2). When these two LLM models were provided the exact same prompt as humans, the percentage of green rulings declined to 35% in the ChatGPT-4 model, but increased slightly to 43.1% in the Claude model (Table 2). LLM models, it appears, have a greater proclivity to interpreting cases in our sample as green.

We then deployed ChatGPT-4 on the entire case sample of more than 12,615 cases. Since we found it to be less accurate in the human-coded sample, we did not run Claude on the larger sample. Here we observe that GPT-4 categorizes rulings as green in about 35.0% of instances. A variable that overrides the values of this inference with the human-coded outcome for the cases that were coded by humans provides a slightly lower estimate of about 32%. The lower percentage of pro-environmental rulings in the broader sample is likely the inclusion of cases less directly related to air pollution in this expanded sample.[8]

4.1 Comparative Performance of AI and Human Coding

Next, we examine how LLM models compare to human coders. To answer this question, we focus on the common sample of 1906 observations that was coded by all three systems and perform several comparisons: (1) Comparison 1: The cases coded by both humans and ChatGPT-4, each with different prompts as defined in the previous section; and (2) Comparison 2: Cases coded by humans as well as ChatGPT-4 using a common prompt that was originally used in the human sample.

Figures 1 and 2 present the detailed case classifications of the common sample for ChatGPT-4 and Claude, respectively. Each chart categorizes cases into four groups: instances both systems identified as environmentally favorable ("green"), cases both deemed not green, and instances where the systems' assessments differed. Complementary confusion matrices are provided in Tables 3 and 4 for ChatGPT-4 and Claude.

Comparison of Figures 1 and the corresponding confusion matrices reveals a significantly stronger alignment between human and AI coding for ChatGPT-4 (Figure 1) compared to Claude. In particular, ChatGPT-4 demonstrates optimal performance when both systems use the same prompt, with improvements primarily driven by improved alignment in classifying cases that are not green.

As previously observed, ChatGPT-4 consistently exhibits a higher tendency to classify cases as environmentally favorable compared to human coders. This systematic bias persists even with improved accuracy, indicating fundamental differences in how AI and human experts interpret environmental impact.

[8]Recall that the human-coded sample, initially focused on air pollution cases before expanding to related environmental issues, is intentionally biased towards air pollution rather than being representative of all environmental cases.

4.2 Additional Checks for Accuracy

TTable 5 presents comparative accuracy statistics for both LLM models against human coding of Indian environmental court cases, analyzing predictions generated using the original human prompt and an improved prompt. The table shows precision, recall, F1 score, and overall accuracy for both prompts. Precision measures how many of the LLM model's "green verdict" predictions were correct, while recall measures how many of the actual green verdicts ChatGPT-4 identified correctly.[9]

We focus first on the results of ChatGPT-4 (first two columns of Table 5). We see that relative to the improved prompt, the use of the human prompt increased precision (from 0.415 to 0.488), but decreased recall (from 0.802 to 0.685). This led to a slight increase in the F1 score (0.547 to 0.570) and overall accuracy (from 0.666 to 0.739). This suggests that when ChatGPT-4 was given the human prompt, it was better able to mimic humans, largely by reducing the number of false positives (541 with the improved prompt and 345 with the human prompt).

Accuracy statistics for Claude are also presented in Table 5. Again, we note that the estimates of precision, recall, F1 Scores, Accuracy and Krippendorff's Alpha are quite similar as the ChatGPT-4 model, but ChatGPT-4 is the more accurate of the two.

4.3 Robustness in Sub-Samples

Finally, we examine the accuracy of ChatGPT-4 as well as Claude 3.5 in subsamples of our data. Panel (a) of Table 6 provides a robustness check of this comparison in a variety of sub-samples that include later years, cases where we successfully defined the petitioners, respondents and judges, cases longer than 300 words (suggesting that they were not procedural cases), air pollution cases, cases at the Supreme Court and NGT, cases in the Delhi NCR region and cases that feature parties other than the pollution control board. All analyses are run on the common sample of cases with the same prompt that was provided to the LLM models as well as humans. The results of both LLM models are presented.

We note that the accuracy ranges from about 75% to 84% with ChatGPT, with the highest accuracy (83. 23%) for cases that do not have action by the Pollution Control Board. Claude is less accurate in each of the sub-samples.

In panel (b) we compare the accuracy of Claude relative to ChatGPT4 and here we find that as expected, Claude's predictions align with ChatGPT4 between 68–74% of the time, with the greatest alignment (89.30%) in the set of cases that is heard at the Supreme Court of the National Green Tribunal of India.

[9] *Precision* is defined as the ratio of correctly predicted (by LLM models) green verdicts to the total predicted green verdicts. *Recall* is defined as the ratio of correctly predicted green verdicts to all verdicts. The *F1 score* is defined as the weighted average (harmonic mean) of precision and recall. *Accuracy* is defined as the ratio of correctly predicted verdicts (both green and non-green) to all verdicts.

5 Discussion

One key observation from our analysis is the higher number of cases coded as green by the LLM models compared to human coders. This discrepancy between the two types of estimates may stem from an unintended bias in the human coding process, arising from the phrasing of our question. The guidance provided to human coders for determining a "positive impact" lacked specificity regarding both a baseline and a timeframe, leading to subjective interpretations. For example, in the case of Kanoon_id 20982084, both human coders had erroneously coded the case as having no impact because the court had rejected a polluting plaintiff's plea to conduct an environmentally harmful activity (use of an illegal thresher machine) and dismissed the case. The order clearly prevented pollution, but the human coders felt that since the baseline levels of pollution were likely to persist, the case would have no impact. The prompt we used with ChatGPT-4 was much more specific about the timeline on which impact needs to be measured, and it also allowed the algorithm to code dismissals as having an environmentally favorable impact.

A second type of bias in the human sample is cynicism about the role of regulatory authorities. When a judgment required the regulator to take an additional action (such as a further test of air or soil samples or conducting additional inspections), the human coders erroneously coded the case as having no positive impact. This is largely because they believed that these actions were unlikely to have a significant impact on actual pollution levels and were merely bureaucratic decisions. However, for ChatGPT-4, any action taken by regulators to reduce pollution, even if it was just further testing, was positively coded.

Finally, some bias may be driven by ChatGPT-4 itself. Recent studies have indicated that ChatGPT-4 exhibits a tendency towards left-leaning viewpoints (Rozado 2023; Motoki, Pinho Neto, and Rodrigues 2024). However, the relationship between political ideology and the interpretation of environmental court decisions is complex. While environmental protection is often associated with progressive political stances, left-leaning perspectives typically express skepticism about the effectiveness of judicial interventions in environmental regulation. This ideological disposition might actually lead human coders with left-leaning views to be more cynical about courts' ability to effectively regulate environmental polluters, potentially resulting in a more conservative assessment of "green" verdicts. Therefore, any systematic differences between ChatGPT-4 and human coders may reflect this nuanced interplay between political ideology and beliefs about institutional effectiveness, rather than a simple bias toward pro-environmental interpretations. This is an interesting area for deeper investigation.

However, in general, our analysis reveals that ChatGPT-4 demonstrates a remarkable ability to approximate human coding of environmental court rulings, achieving accuracy rates above 70%. However, this high performance is nuanced by a consistent tendency of the AI to classify cases as environmentally favorable ("green") more frequently than human coders.

These findings highlight both the potential and the limitations of using large-language models in legal analysis. Although ChatGPT-4 shows promise in automating the coding of

environmental court rulings, its tendency to overestimate pro-environmental outcomes underscores the continued importance of human oversight and the need for careful calibration of AI tools in legal research contexts.

6 Conclusion

This study demonstrates the potential of AI, specifically ChatGPT-4, to improve the analysis of judicial environmental court rulings in India. We show that AI can effectively summarize complex legal documents with high accuracy, achieving 73% agreement with human coders on a sample that featured a common prompt. Our creation of a new comprehensive dataset of 12,615 environmental cases opens new avenues for research in environmental policy and judicial decision-making in India. These findings also suggest that AI tools can significantly improve the efficiency and scalability of legal research well beyond the Indian context, particularly in cases where administrative data are not yet standardized. Future research could explore the application of these methods to other areas of law and policy, potentially revolutionizing the way we analyze large-scale legal and administrative datasets.

References

Aher, Gati V, Rosa I Arriaga, and Adam Tauman Kalai (2023). "Using large language models to simulate multiple humans and replicate human subject studies". In: *International Conference on Machine Learning*. PMLR, pp. 337–371.

Ash, Elliott et al. (2022). "Measuring gender and religious bias in the Indian judiciary". In.

Athey, Susan and Guido W Imbens (2019). "Machine learning methods that economists should know about". In: *Annual Review of Economics* 11, pp. 685–725.

Bhupatiraju, Sandeep, Daniel L Chen, Shareen Joshi, and Peter Neis (2024). "Impact of free legal search on rule of law: Evidence from Indian Kanoon". In.

Bhupatiraju, Sandeep, Daniel L Chen, Shareen Joshi, Peter Neis, and Shashank Singh (2024). "Litigation as Scrutiny: A Four Decade Analysis of Environmental Justice, Firms, and Pollution in India". In.

Bhupatiraju, Sandeep, Daniel Li Chen, and Shareen Joshi (2021). "The Promise of Machine Learning for the Courts of India". In: *Nat'l L. Sch. India Rev.* 33, p. 463.

Brown, Tom et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Chandra, Aparna, William HJ Hubbard, and Sital Kalantry (2017). "The supreme court of India: a people's court?" In: *Indian Law Review* 1.2, pp. 145–181.

Damle, Devendra and Tushar Anand (2020). "Problems with the e-Courts data". In: *National Institute of Public Finance and Policy Working Paper* 314.

Do, Quy-Toan, Shareen Joshi, and Samuel Stolper (2018). "Can environmental policy reduce infant mortality? Evidence from the Ganga Pollution Cases". In: *Journal of Development Economics* 133, pp. 306–325.

Einav, Liran and Jonathan Levin (2014). "Economics in the age of big data". In: *Science* 346.6210, p. 1243089.

Geiger, Christian P and Jörn von Lucke (2011). "Open Government Data[1] Free accessible data of the public sector". In: *CeDEM11*, p. 183.

Ghosh, Shibani, ed. (2019). *Indian Environmental Law: Key Concepts and Principles*. Orient BlackSwan.

Haskins, Ron (2018). *Evidence-based policy: The movement, the goals, the issues, the promise*.

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

Korinek, Anton (2023). "Generative AI for economic research: Use cases and implications for economists". In: *Journal of Economic Literature* 61.4, pp. 1281–1317.

Liang, Yingshuai (2023). "Exploring the Path of Judicial Big Data to Enhance Data Governance Capability". In: *Applied Mathematics and Nonlinear Sciences* 9.1.

Moraes, Beatriz Fruet de, Fabrício Castagna Lunardi, and Pedro Miguel Alves Ribeiro Correia (2024). "Digital Access to Judicial Services in the Brazilian Amazon: Barriers and Potential". In: *Social Sciences* 13.2, p. 113.

Motoki, Fabio, Valdemar Pinho Neto, and Victor Rodrigues (2024). "More human than human: measuring ChatGPT political bias". In: *Public Choice* 198.1, pp. 3–23.

Penner, Andrew M and Kenneth A Dodge (2019). "Using administrative data for social science and policy". In: *RSF: The Russell Sage Foundation Journal of the Social Sciences* 5.2, pp. 1–18.

Posner, Richard A (2014). *Economic analysis of law*. Aspen Publishing.

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

Reiling, Dory and Francesco Contini (2022). "E-justice platforms: Challenges for judicial governance". In: *IJCA*. Vol. 13. HeinOnline, p. 1.

Rozado, David (2023). "The political biases of chatgpt". In: *Social Sciences* 12.3, p. 148.

Savelka, Jaromir et al. (2023). "Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise?" In: *arXiv preprint arXiv:2306.13906*.

Susskind, Richard (2019). *Online courts and the future of justice*. Oxford University Press.

Ziems, Caleb et al. (2024). "Can large language models transform computational social science?" In: *Computational Linguistics* 50.1, pp. 237–291.

Tables and Figures

Table 1: Acts Cited in Our Database

Act	Number of cases
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
Negotiable Instruments Act, 1881	345
Article 14 in Constitution of India	336
Article 227 in Constitution of India	329
Indian Evidence Act, 1872	314
Air Force Act, 1950	280
Article 48a in Constitution of India	243
Mines and Minerals (development and regulation) Act, 1957	215

Notes: We include here any acts that were cited at least 300 times. Additional noteworthy acts that are not included here include: additional acts from the Indian constitution (32, 51g), the Land Acquisition Act of 1894, various state forest acts, the Prevention of Corruption Act of 1988, the Income Tax Act of 1961 and the water (prevention and control of pollution) Cess Act of 1977. There were numerous instances of landmark verdicts (such as MC Mehta versus Union of India cases) that were also cited. A full list is available upon request.

Table 2: Summary Statistics of human sample and expanded sample

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
Number of human readers	1,702	1.440	0.500	1.0	3.0
Sum of scores of human readers	1,702	0.371	0.561	0.0	3.0
Appeal case (Human coding)	1,702	0.283	0.450	0.0	1.0
Constitutional case (Human coding)	1,702	0.127	0.333	0.0	1.0
Government plays a role (Human coding)	1,702	0.000	0.000	0.0	0.0
Case Relevant to the Environment (Scale 0-2)	1,904	0.830	0.376	0.0	1.0
PCB Action (GPT4)	1,910	0.472	0.499	0.0	1.0
Regulator Action (GPT4)	1,910	0.564	0.496	0.0	1.0
Length of case (characters)	1,910	4915	11705	131	351450
Delhi NCR Region	1,910	0.282	0.450	0.0	1.0

Expanded sample (Coded by ChatGPT-4)	N	Mean	SD	Min	Max
Green Verdict (GPT4 coding)	12,607	0.350	0.478	0.0	2.0
Green Verdict (human coding)	12,613	0.314	0.465	0.0	2.0
Order	12,615	0.199	0.400	0.0	1.0
Regulator Action (GPT4)	12,615	0.357	0.479	0.0	1.0
PCB Action (GPT4 coding)	12,615	0.273	0.446	0.0	1.0
Politician Action (GPT4 coding)	12,615	0.045	0.207	0.0	1.0
Number of petitioners	12,615	2.119	6.084	0.0	250.0
Number of respondents	12,615	3.102	6.256	0.0	148.0
Number of judges	12,615	1.534	0.918	0.0	7.0
Number of states	12,615	1.065	0.941	0.0	35.0
Supreme Court case (GPT4 coding)	12,615	0.032	0.177	0.0	1.0
High Court case (GPT4 coding)	12,615	0.689	0.463	0.0	1.0
NGT case (GPT4 coding)	12,615	0.226	0.418	0.0	1.0
Delhi NCR Region	12,615	0.290	0.454	0.0	1.0

Notes: (i) “Green Verdict" is a binary variable that takes 1 if the average human coder regarded the court order or judgement as pro-environment (and 0 otherwise); (ii) “Order" takes value 1 if the ruling is an interim ruling (and 0 otherwise); (iii) “Regulator Action (GPT4 coding)" and “PCB Action (GPT4 coding)" take value 1 if the order involved any action from a regulatory body or specifically the Pollution Control Board (PCB); (iv) “Supreme Court Case", “High Court case" and “NGT case" take value 1 if the case is heard at the Supreme Court, High Court or National Green Tribunal respectively (and 0 otherwise); (v) Delhi NCR is a dummy variable that takes value 1 if the case stems from the broader Delhi National Capital Region area (and 0 otherwise).

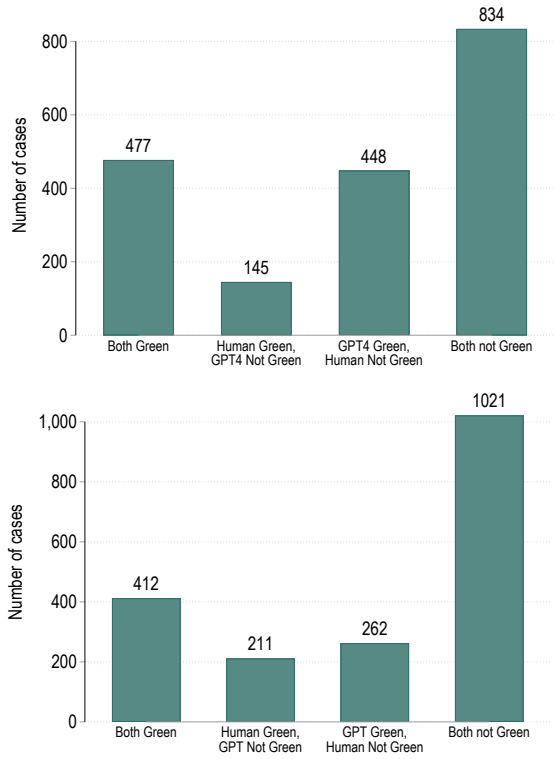

Figure 1: Comparison of ChatGPT-4 with humans. Panel (a): improved prompt; Panel (b): common prompt.

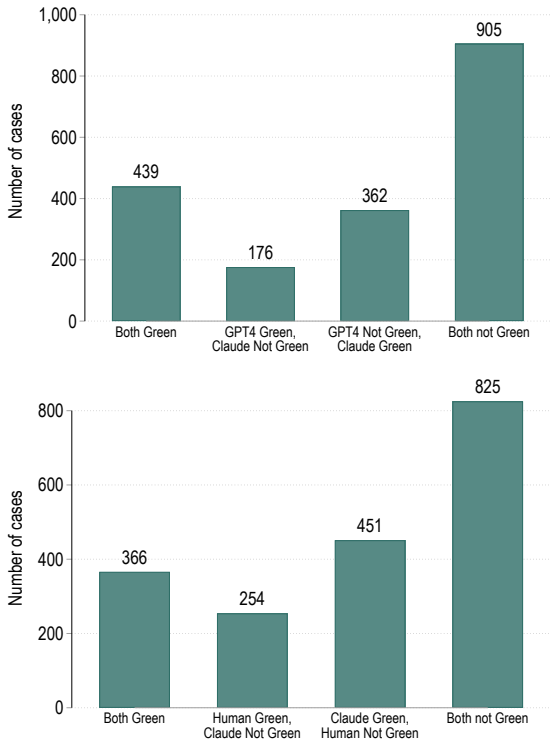

Figure 2: Comparison of Claude with humans. Panel (a): improved prompt; Panel (b): common prompt.

Table 3: Confusion Matrix – ChatGPT-4

		ChatGPT-4 with improved prompt				**ChatGPT-4 with human prompt**		
		0	1	Total		0	1	Total
Human	0	884	541	1425	0	1081	345	1426
	1	95	384	479	1	151	329	480
	Total	979	925	1904	Total	1232	674	1906

Notes: This confusion matrix compares ChatGPT-4's prediction of green cases to the human sample (which we here regard as the true sample). This matrix tabulates the number of true negatives (both 0), true positives (both 1), false positives (GPT4 is 1 and human is 0), and false negatives (GPT4 is 0 and human is 1).

Table 4: Confusion Matrix – Claude

		Claude with improved prompt				**Claude with human prompt**		
		0	1	Total		0	1	Total
Human	0	905	362	1267	0	825	451	1276
	1	176	439	615	1	254	366	620
	Total	1081	801	1882	Total	1079	817	1896

Notes: This confusion matrix compares Claude's prediction of green cases to the human sample (which we here regard as the true sample). This matrix tabulates the number of true negatives (both 0), true positives (both 1), false positives (GPT4 is 1 and human is 0), and false negatives (GPT4 is 0 and human is 1).

Table 5: Additional Statistics for Accuracy

	GPT4		Claude	
Prompt Type:	**Improved**	**Human**	**Improved**	**Human**
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
Accuracy	0.666	0.739	0.714	0.629
Krippendorff's Alpha	0.282	0.383	0.392	0.210

Notes: *Precision* is defined as the ratio of correctly predicted (by ChatGPT) green verdicts to the total predicted green verdicts. *Recall* is defined as the ratio of correctly predicted green verdicts to all verdicts. The *F1 score* is defined as the weighted average (harmonic mean) of precision and recall. *Accuracy* is defined as the ratio of correctly predicted verdicts (both green and non-green) to all verdicts.

Table 6: Accuracy in Sub-Samples

Sub-Samples	**N**	**Accuracy GPT4 vs. Humans**	**N**	**Accuracy Claude vs. Humans**
Panel (a): Common cases, LLM models compared with human prompt				
All cases in this sample	1906	75.18%	1896	62.82%
Cases after 1990	1906	75.18%	1896	62.82%
Cases with 1+ petitioner, respondent and judge	1880	75.21%	1870	62.78%
Cases that are greater than 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 PCB	1002	83.23%	996	66.16%

		GPT4 vs. Claude
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 at least 1 petitioner, respondent and judge	1870	68.50%
Cases that are greater than 300 words	1790	67.37%
Cases that are regarded as being 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 (broadly defined)	538	67.47%
Cases featuring no action by Pollution Control Board	996	71.39%

Appendix

6.1 Questions for ChatGPT3.5

- Q1: In the following summary of a judgment order, predict on a scale of 0-100 if the order had a pro-environmental impact where 100 is the most pro-environment. Give a number from 0 to 100 and then give the reason separated from the binary answer with a colon.
- Q2: In the following judgment order summary, did the court ask the pollution control board to take an action? Give a binary answer, yes or no and then describe the action separated from the binary answer with a colon.
- Q3: In the following judgment order summary, did the court tell the respondent to follow the law, or did the court tweak the law in its judgment? Give a binary answer, 'follow' or 'tweak' and then give the reason separated from the binary answer with a colon.
- Q4: In the following judgment order summary, did the court's decision compel action by a politician or regulator? Give a binary answer, yes or no and then describe the action separated from the binary answer with a colon.
- Q5: In the following judgment order summary, predict on a scale of 0-100 if the case is relevant to the environment where 100 is the most environmentally relevant. Give a number from 0 to 100 and then give the reason separated from the binary answer with a colon.
- Q5 prompted to identify if the case was relevant to environment, which can help exclude junk (i.e cases not relevant to environment)

6.2 Questions for ChatGPT-4.0

- Summarize the final decision of the order in at most 200 words
- Extract the outcome 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
- Did the court ask the pollution control board to take action? Give a binary answer, yes or no and do not write anything else.
- Did the court tell the respondent to follow the law, or did the court tweak the law in its judgment? Give a binary answer, 'follow' or 'tweak' and do not write anything else.

- Did the court's decision compel action by a politician or the Government? Give a binary answer, yes or no and do not write anything else.
- Did the court's decision compel action by a regulatory authority (apart from the Government)? Give a binary answer, yes or no and do not write anything else.
- Is the underlying dispute regarding air pollution? Give a binary answer, yes or no, and do not write anything else.
- Extract names of all the petitioners.
- Extract names of all the respondents.
- Please extract the location of the jurisdiction of the case. Please note that the location of the court might or might not be the same as that of the jurisdiction. For example, if the case is heard in high court at Allahabad but the underlying conflict lies in Lucknow, the jurisdiction should be Lucknow and not Allahabad. Please give the name of the district, state and the city. If the jurisdiction is the entire state or country, write the name of the state and/or country. If there are multiple jurisdictions include them in the list
- Extract the case numbers from the text. They are usually found within the first few lines where the petitioners and respondents are listed. They are typically of the format "<case_type with some_text> no. <numbers> / <year>". For example: "S.B. Civil Writ Petition No. 679 2018" and "Original Application No. 10512018" They may not always strictly adhere to the specified format i.e they can be slightly different for ex: "OWP No. 106 of 2017" , "MP No. 01 of 2017" etc. Note that some orders may contain multiple cases and therefore have multiple case numbers.

CASE CODING MANUAL

A guide for coding and classifying judgments from Indian courts

Data and Evidence for Judicial Reform (DE JURE)
Development Impact Evaluation (DIME)
The World Bank

Last updated: March 2021

TABLE OF CONTENTS

Introduction

At the DE JURE program, we aim to harness the potential of recent changes in data availability to expand the evidence base on the economics of justice reform through rigorous analysis and experimentation.

A large amount of judicial data has become available in India since the launch of the e-courts national portal in 2013. Our team of researchers has been using this data to conduct in-depth research on various aspects of the judiciary and its impact on economic outcomes.

As part of this endeavor, we are also studying the impact of environmental litigation in India on actual environmental outcomes such as air pollution levels, water quality, etc. In this regard, we require you to read and extract relevant information from a dataset of environmental judgments passed by Indian courts. Your input will enable us to categorize cases and record the characteristics of these judgments that are relevant to the studies we conduct.

This document provides useful instructions and guidance on how to navigate the case coding portal and review judgments to extract relevant information.

We thank you for your valuable assistance with our research endeavors!

Accessing the Form

Step 1: Visit https://airpollutioncases.herokuapp.com/room/airpollution/.

Step 2: Enter the participant label assigned to you.

Your participant label will be sent to you separately by email.

Step 3: You will be directed to a page with the case to be coded and the relevant fields to be filled out by you.

In case of any questions or technical difficulties, please reach out on the Slack channel or write to ***Ritesh Das*** *at* ***rdas4@worldbank.org***.

Reading the Judgments

The key to this exercise is to read the judgments carefully, with the aim to extract relevant information. Some information like case title and judge names are easy to identify. However, some fields are more complicated and completing them accurately will require you to have a deeper understanding of the facts of the case, arguments made in court, and the final verdict.

While there is no uniform format for judgments passed by the courts, all judgments are likely to have the following components:

Cause Title:

The cause title contains the name of the court, case number, and party names. Located at the very top of the judgment it usually follows the following format:

IN THE HIGH COURT OF KARNATAKA

Civil Writ Petition No. 1 of 2021

John Doe (Petitioner)

Verus

State (Respondent)

At times, the names of the judge(s) will be included in the cause title as "*Before: Hon'ble Mr./Ms. Justice X*". Other times this may be mentioned after the cause title.

You may also come across judgments where two or more cases have been heard together as tagged matters. This usually happens if two or more cases deal with the same cause of action or have similar prayers for relief.

Facts:

The main text of the judgment usually begins with a summary of the facts of the case. This section is very useful to understand case context and extract information regarding parties, advocates, districts, etc.

Arguments:

The facts will generally be followed by a brief record of the arguments made before the court by the respective advocates.

Legal Discussion:

A section of the judgment is often dedicated to summarizing the relevant legal provisions and precedents relevant to the facts of the case. In this section, the judges may also discuss the history or evolution of the law, their opinion of the legal position on the subject, etc.

Final Orders:

The final orders are the actionable part of the judgment and set out the final decision (dismissed, appeal allowed, relief granted, etc.). This is usually set out at the very end of the judgment.

It is important to note that this is only a general guide on the conventional format of a judgment. However, each of these components will not always be distinctly identifiable in all the judgments you encounter.

Filling Out the Form

Case ID

- The case ID refers to the case number which appears at the beginning of the judgment - either before or as part of the cause title. The case ID will generally be in an abbreviated format. For example, WP No. 1 of 2021, CA No. 2 of 2021 etc.

Judge Names

- Judge names are often written under the "*Coram*" section of the judgment, which usually appears right after the cause title.
- The judge names may also be included in the cause title as "*Before: Hon'ble Mr./Ms. Justice X*"
- Check to see if it is a single bench or division bench (two judges). In some cases, there may also be 3 or more judges adjudicating on the case.
- Start typing the name of the judge. This should prompt a drop-down menu with judge names to appear on your screen. You can then select the name of the judge relevant to the judgment you are coding.
- In case the judge's name does not appear in the drop-down menu, you can also fill it manually by typing in the name. However, while typing in the name, please remember to only include their first name, last name, and middle name (if any). Do not include prefixes such as "*Hon'ble Mr. Justice*".

Petitioner and Respondent Names

- Check to see the number of petitioners and respondents in every judgment.
    - The suffix "& Anr." after a petitioner/respondent name in the cause title implies there are two petitioners/respondents.
    - The suffix "& Ors." after a petitioner/respondent name in the cause title implies there are more than two petitioners/respondents.
- In some cases, the names of all petitioners and respondents will be included in the cause title. However, if all party names are not mentioned in the cause title, check the judgment text. Party names are often included in the judgment text as part of the facts and arguments.
- **Please include only one petitioner or respondent name in a single field**. If there is more than one petitioner or respondent, you can add fields by clicking on the "**+**" button.

Advocates

- Advocate names will typically appear after the cause title along with phrases like “Present” or “On behalf of”.
- When filling in advocate names, only include their first name, last name, and middle name (if any). Do not include prefixes such as “*Adv*”, “*Sr. Adv*”, “*Mr.*”, “*Ms.*”, etc.
- **Please include only one advocate name in a single field**. If there is more than one advocate for the petitioner(s) or respondent(s), you can add fields by clicking on the “**+**” button.

Company Involved

- For this field, you need to identify if any company is a party to the litigation, either as a petitioner or respondent. If you find that a company is a party, please enter the name of the company in this field.
- Please keep the following in mind while filling out this field:
    - Party names suffixed with “*& Co.*”, “*Ltd.*”, “*Pvt. Ltd.*” etc. are companies.
    - “*M/s.*” is often used as a prefix while recording company names in judgments. Although technically the prefix “*M/s.*” should only be used for firm names, it is a practice followed by Indian courts to use it as a prefix for company names as well.

District

- This field refers to the district in which the cause of action arises. For example, in cases of water pollution, you will fill in the name of the district where the water body in question is located.
- To fill out this field, start typing the name of the district. This should prompt a drop-down menu with district names to appear on your screen. You can then select the name of the relevant district.
- In case the district name does not appear in the drop-down menu, check to see if there is a similar district name spelled slightly differently. Sometimes the name of the district may have been misspelled. If you still cannot find the district in the drop-down list, you can simply type in the name of the district.
- The name of the district is not always evident from the judgment and there may be judgments where the district is not mentioned at all. In such a situation, you can leave this field blank. However, be sure to examine the judgment text (especially the facts component) carefully for this information.

- The district can sometimes be determined through other details within the judgment text. For example, details about the District Magistrate or Police Station. However, be sure to read the facts carefully before concluding the district

State

- This field refers to the state in which the cause of action arises.
- To fill out this field, start typing the name of the state. This should prompt a drop-down menu with state names to appear on your screen. You can then select the name of the relevant state from this list.
- It is safe to assume that if a particular state high court is hearing the case, the issue being contested has arisen within that state. This is what gives the court the jurisdiction to hear the matter in the first place. If this is not the case, and the issue being contested relates to a different state, an explanation will usually be provided in the judgment text. However, this is rare and will happen only in exceptional circumstances.
- In the case of National Green Tribunal or Supreme Court judgments, the state should be easily identifiable from the facts.

Appeal Case

- In this field, we want to determine whether the case was brought before the court as an appeal or not.
- You can determine this by:
    - *The case number* - Each court has slightly varying nomenclatures for different kinds of appeals. Appeals will usually be categorized as civil appeal, criminal appeal, etc. They may also be abbreviated as C.A. or Crl.A.
    - *Facts* - The facts stated in the text of the judgment will typically mention that a case has been filed as an appeal.
- If you determine that the case is an appeal, select “Yes” from the drop-down menu. Else, select “No”.

Constitutional Case

- Constitutional cases are those where **a substantial question regarding the interpretation of a Constitutional provision** is involved.
- **A simple mention of the Constitution in the judgment text is not enough**. For example, many of the cases you come across will be writ petitions and may mention that they have been filed under Article 226/227 of the Constitution. This, by itself, is

insufficient to categorize the case as one that is Constitutional. However, if for example, a case is discussing whether the right to clean air/water is a fundamental right under the Constitution, it should be categorized as "Constitutional" because it involves interpretation of a Constitutional provision.

- If you determine that the case is a Constitutional case, select "Yes" from the drop-down menu. Else, select "No".

Government's Role

- In this field, we want to determine if the Government is a party to the litigation.
- Here, "Government" includes:
    - The Central Government (Union of India) or any of its Ministries and/or Departments;
    - The State Government (State of Punjab, State of Karnataka, etc.) or any of its Ministries and/or Departments;
    - Local authorities (Municipalities, etc.); and
    - Statutory bodies.
- Select whether the Government is a petitioner or respondent from the drop-down menu. In case the Government is involved both as a petitioner and respondent (as in the case of inter-state or inter-departmental disputes) select "Both". Select "None" if the Government is not involved in the litigation.

Environmental Impact

- In this field, we want to determine whether the judgment is likely to have a positive impact on the environment or not.
- 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 for being heard afresh without passing any orders on the merits of the case, will also fall in this category.