

BEYOND PATTERN MATCHING: USING LARGE LANGUAGE MODELS TO UNLOCK COURT DATA *

Chetana Sabnis

June 2026

This article examines the use of large language models (LLMs) to collect and classify information in court data. While computational approaches have traditionally relied on pattern matching techniques, these methods struggle to obtain detailed information from court data. Drawing on projects involving more than 2,700 Indian court cases and 4,800 U.S. state cases, I argue that LLMs offer three advantages over traditional approaches: they can identify concepts from natural language descriptions, synthesize information across complex documents, and lower the technical barriers to large-scale text analysis. I develop a framework to use LLMs effectively so that the model draws on relevant information and produces reliable outputs. In a hand-coding validation of 600 cases, the framework produced classifications that mostly aligned with human coding. The article concludes by discussing the ethical considerations, including environmental costs, research labor implications, and risks of concentrating power within a small number of firms.

*Under review. Please do not cite or circulate without permission. Comments welcome at chetana.sabnis@yale.edu.

1. INTRODUCTION

Court cases are among the richest data sources available to political scientists. A single opinion contains a factual narrative, competing legal arguments, judicial reasoning, and a disposition. However, despite a growing “text as data” movement in the discipline (Grimmer and Stewart 2013; Lucas et al. 2015; Liebman et al. 2017), computational approaches have typically extracted only surface-level features from these cases.

The dominant approach to collecting and classifying court data is based solely on pattern matching (Lazarev 2019; D. D. Choi, Harris, and Shen-Bayh 2022; Bhupatiraju et al. 2024; Joshi et al. 2024; Ash et al. 2025; Truscott and Romano 2026). Using techniques such as string matching, regular expressions, keyword dictionaries, character n-grams, and word embeddings, pattern matching identifies concepts through predefined textual markers. While effective for extracting clearly specified information, it is less useful when extracting information embedded within detailed factual narratives, competing legal arguments, and judicial reasoning.

Large language models (LLMs) provide a compelling alternative. LLMs are artificial neural networks “trained on a large portion of the internet” (Somers 2025). This training allows them to identify complex patterns in language that correspond to underlying social, cultural, and political concepts. These associations serve as a “proxy for semantic meaning” (Pavlus 2024). As a result, these models can extract *and* analyze information from a text that would be difficult to identify through pattern matching alone.

In the context of court data, I find that LLMs offer three distinct advantages over a pure pattern matching approach: First, they **identify concepts from natural language descriptions** without requiring extensive hand-crafted dictionaries, training data, and rules. Second, they can **synthesize and classify information distributed across long and complex text** rather than relying on localized textual features. Third, they **substantially lower the technical barriers to advanced text analysis** given they can be deployed through commercially available APIs using natural language instructions.

Recent work has shown that LLMs including those that are off-the-shelf like OpenAI’s GPT-4 or GPT-4-turbo can annotate and classify political text rivaling or exceeding human coders (Gilardi, Alizadeh, and Kubli 2023). This holds true for court data as well, which tends to be long and complex and involves information that is more subtly expressed (J. H. Choi 2024; Deroy,

K. Ghosh, and S. Ghosh 2024; Heddaya et al. 2024). Yet these studies also highlight the risks of hallucinations and biases, making human oversight and validation essential.

Despite growing interest in LLMs, many political scientists remain uncertain about how to use them effectively in judicial research. In a recent conversation, a graduate student described uploading a lengthy court opinion into GPT-4 only to receive outputs that were inaccurate. While LLMs are powerful tools for collecting and classifying information from legal texts, their performance depends heavily on how they are used.

Based on applying an “LLM-with-human-in-the-loop” approach in my own research, which involved collecting and classifying more than 2,700 Indian High Court cases and 4,800 U.S. state appellate cases, I propose a two-part framework for using LLMs on court data so that the LLM draws on the right information and produces reliable outputs. To ensure relevance, the first part concerns structuring the case document and prompt so that the LLM can identify and analyze the information most pertinent to the question at hand. To ensure reliability, the second part concerns constraining the model and building in checks to minimize hallucinations and biases. I validate the framework by comparing a hand-coded classification of a random sample of 600 cases against the model’s output. The two outputs aligned on 95% of the classifications.

The article also considers ethical tradeoffs associated with LLM use, including environmental costs, loss of opportunities for student RAs, and the risk of furthering the concentration of power in a small number of firms that control the development and deployment of LLMs. These considerations should factor into decisions about whether LLMs are preferable to alternative approaches.

The article proceeds as follows. First, I explain why LLMs were better suited than pattern matching alone for my research objectives. I then present the two-part framework, illustrate how it improves the effectiveness of LLMs for legal text analysis, and discuss validating the framework using a hand-coded sample of 600 cases. I conclude by discussing the ethical considerations that should guide the use of LLMs in political science research. This article contributes to a growing literature on the rigorous and theoretically informed integration of LLMs into political science research (Lee et al. 2025; Weidmann, Faulborn, and García 2026).

2. WHAT PATTERN MATCHING CANNOT DO

The limitations of pattern matching become clearer when considering the requirements of my own research. I sought to identify individuals who lacked default recognition as family but nonetheless claimed that status in court. In India, this meant women in extramarital relationships who sought recognition as wives. In the U.S., it meant individuals who were neither biological nor legal parents but sought recognition as parents on the basis of kinship and caregiving. Beyond identifying these cases, I also wanted to collect detailed information about relationship characteristics discussed by courts, such as whether a woman had been deceived into marrying a man who was already married or whether a claimant had housed and cared for the child. Collecting and classifying this information at scale through conventional approaches would have required substantial technical expertise, labeled training data, and custom model training, increasing both the budget and time for this project. In this context, LLMs proved a compelling alternative.

First, LLMs allowed me to identify concepts using natural language descriptions. This was important because the same concept appeared in highly variable language. In the India dataset, participation in a marital ritual could be described as a marriage being “solemnized at Jaynati Mandir,” a woman wearing “‘Sindoor’,” or a couple having “exchanged garlands.” Similarly, in the U.S. dataset, caregiving could be described through activities such as “raising the children in her home,” “enrolling them in gymnastics and dance classes,” and “attending parent-teacher conferences.” Capturing these varied descriptions through pattern matching approaches would require extensive dictionaries and hand-crafted rules. By contrast, LLMs allowed me to specify the underlying concept— “marital ritual” for the Indian dataset, “caregiving” for the U.S. dataset—and identify it across diverse linguistic expressions.

Second, LLMs enabled inference across long and complex case transcripts. In the India project, the decision to include a case in the dataset depended on a combination of facts: a woman claiming recognition as a wife, the existence of a prior marriage, and a dispute over spousal entitlements. No single phrase reliably identified such cases. Instead, determining whether a case belonged in the dataset required drawing inferences based on reading the full case and assembling the “right” set of facts. The same challenge arose when classifying outcomes. Existing approaches often infer outcomes from whether a petition was “allowed” or “dismissed” (Bhupatiraju et al. 2024). In my project, because women in disputed relationships could appear as either petitioners

or respondents, identifying whether a judgment benefitted them required understanding the parties, the claims, and the court’s reasoning before interpreting the final disposition. Such inference is difficult to achieve through conventional pattern matching but is possible with LLMs.

Finally, LLMs lowered the cost and technical barriers to large-scale text analysis. Whereas advanced pattern matching approaches often require more budget and time, commercially available LLMs can be deployed through natural language instructions and relatively affordable APIs. This not only reduced the resources required for the project but also made it easier to revise coding schemes as my understanding of the data evolved. In other words, I could easily revise the prompt to adjust the classification criteria without worrying about rebuilding an entire analytical pipeline. Ultimately, using LLMs enabled screening more than 10,000 case transcripts in each legal context in order to construct the final datasets.²

3. HOW TO USE LLMs FOR COURT DATA

The workflow presented here assumes that the researcher has some preliminary understanding of programming especially if they are looking to collect and classify information across a wide sample of case transcripts. In my research, I used Python to automate retrieving more than ten thousand case transcripts from legal databases³, performing preliminary pattern matching to ensure that the obviously irrelevant cases were filtered out, passing relevant cases to the LLM, and compiling outputs into structured datasets.⁴

While calling the LLM was one of many components in my code, it was central for collecting and classifying all the data. As mentioned, LLMs were used based on a two-part framework. The first part concerned feeding the most relevant portions of the case document and designing the prompt in a manner that ensured the right information was analyzed for the specific task. The second part concerned constraining the model and building in checks to minimize hallucinations

²Based on an estimate of \$0.03 per 1,000 tokens, reading and generating outputs across 20,000 transcripts with about 10 pages each, which would mean a total of 120-180 million tokens, roughly turned out to be about \$4,500. Tokens are essentially the basic units of language for AI. One token can include spaces, punctuation marks, and/or a couple of characters.

³Indian High Court cases were collected through a set of systematic search queries from Indian Kanoon, an online public legal repository. U.S. state court cases were available on Nexis but unlike Indian Kanoon, Nexis did not allow me to directly have the LLM read cases off the websites; I therefore downloaded every case matching my search criteria and had my code pass these cases that were stored locally to the LLM.

⁴This is technically cumbersome but it is much less demanding than traditional pattern-matching approaches, and LLMs can assist in generating and refining code.

and biases and ensure the outputs were reliable. I discuss each part in turn.

3.1. FOCUSING ON RELEVANT INFORMATION

As much as LLMs are capable of processing lengthy case documents, many of these documents are complex, containing procedural histories, factual narratives, competing legal arguments, and judicial reasoning. Of all this information, only some will be relevant to the question at hand. Asking an LLM to read an entire opinion, however specific the question, risks diluting the model's attention. The goal should therefore be to ensure that the LLM focuses on the most relevant information necessary to answer the question.

This principle is reflected in how researchers use a model's context windows (J. H. Choi 2024). A context window is essentially how much text a model can hold in its working memory during a single interaction. Although many of today's LLMs can process well over 100,000 tokens at once, researchers often feed them far less, instructing the model to read no more than 8,000-10,000 tokens at a time.⁵ This reduces costs and processing time but critically, ensures that the model's attention is focused on the "right" information.

Capping tokens, however, does not help with directing the model to the right parts of the document. For this, I found two steps especially helpful: first, leveraging the case document's format; second, designing prompts to align with whether the task calls for direct or indirect case information.

3.1.1. LEVERAGING THE CASE DOCUMENT FORMAT

Many court opinions follow a predictable format that can be exploited to improve classification accuracy. U.S. state appellate opinions, for example, typically contain sections such as case summary, headnotes, factual backgrounds, and discussion. Figure 1 shows a sample of how the case document is formatted. Even before sending documents to the LLM, I used a basic pattern matching function to remove sections irrelevant to the analysis such as "Core Terms" and "Headnotes." This ensured that the model processed only the portions of the opinion most likely to contain relevant information.

⁵The average case transcript is 10 pages, or roughly 8000-10,000 tokens.

Egan v. Fridlund-Horne, 221 Ariz. 229	Page 1 of 16	Egan v. Fridlund-Horne, 221 Ariz. 229	Page 4 of 16
⚠ Caution As of: February 3, 2026 3:49 PM Z			
<u>Egan v. Fridlund-Horne</u> Court of Appeals of Arizona, Division One, Department A April 14, 2009, Filed No. 1 CA-SA 08-0240			
Reporter 221 Ariz. 229 **; 211 P.3d 1213 **; 2009 Ariz. App. LEXIS 56 ***; 554 Ariz. Adv. Rep. 6 MICHELLE ANNE EGAN, Petitioner, v. THE HONORABLE ELAINE FRIDLUND-HORNE, Judge of the SUPERIOR COURT OF THE STATE OF ARIZONA, in and for the County of COCONINO, Respondent Judge, THERESE MARIE HOCHMUTH, Real Party in Interest.	would not be away from her mother more than five consecutive days.		<u>custody</u>, and thus not nearly as invasive of parents' rights. Counsel: For Petitioner: Bryon Middlebrook, Bryon Middlebrook, PC, Flagstaff. For Real Party in Interest: Mikkel Jordahl, Mikkel Jordahl PC, Flagstaff. Judges: MICHAEL J. BROWN, Presiding Judge. CONCURRING: MARGARET H. DOWNIE, Judge. BARKER, Judge, specially concurring. Opinion by: MICHAEL J. BROWN
Prior History: [***1] Petition for Special Action from the Superior Court in Coconino County, Cause No. DO-2008-0205. The Honorable Elaine Fridlund-Horne, Judge. Disposition: JURISDICTION ACCEPTED; RELIEF GRANTED.	Overview The parties were partners in a same-sex relationship, during which the mother gave birth to a child that they raised together for seven years until the relationship broke up. The nonparent sought additional <u>visitation</u> of the child under <u>Ariz. Rev. Stat. § 25-415(C)</u> (2007). On appeal of the trial court's order, the court held that the trial court's awarding the nonparent equal <u>visitation</u> rights was not reasonable and that it failed to accord significant weight to the mother's pre-petition voluntary agreement to allow the nonparent some <u>visitation</u>. Nothing in the record suggested that the trial court applied the presumption that the mother acted in the best interests of her child in determining the amount of <u>visitation</u> the nonparent should receive. The record also did not reflect sufficient evidence to rebut the presumption. There was also no indication that the trial court applied the factors found in <u>Ariz. Rev. Stat. § 25-409(C)</u> (2007). The trial court awarded <u>visitation</u> significantly in excess of what the parties allegedly agreed to in July 2007 without determining whether such agreement was in the child's best interests. Outcome The order was vacated and the case was remanded for further proceedings.	HN175-4: A trial court's consideration of a petition for <u>in loco parentis</u> <u>visitation</u> under <u>Ariz. Rev. Stat. § 25-415(C)</u> (2007) is sound when based on the following procedural and evidentiary safeguards: (1) the court should apply a rebuttable presumption that a fit parent's decision to deny or limit <u>visitation</u> was made in the child's best interests; (2) the court must give "some special weight" to the parent's determination of whether <u>visitation</u> is in the child's best interests and give "significant weight" to the parent's voluntary agreement to permit some <u>visitation</u>; (3) the court must consider the best interests factors listed in <u>Ariz. Rev. Stat. § 25-409(C)</u> (2007); and (4) the court should take into consideration other relevant best interests factors such as the degree to which the parent has consented to and fostered the nonparent's relationship with the child, including any agreements the parties made as to <u>visitation</u> arrangements. Furthermore, the burden of proof is on the nonparent to rebut the presumption that a fit parent's decision to limit or deny <u>visitation</u> is in the child's best interests.	Opinion [**1215] [***231] BROWN, Judge P1 This special action addresses questions of first impression relating to the superior court's decision to grant substantial <u>visitation</u> rights to a person <u>standing in loco parentis</u> under <u>Arizona Revised Statutes ("A.R.S.") section 25-415(C)</u> (2007). Because we conclude that the court failed to employ adequate procedural and evidentiary safeguards to protect the interests of the legal parent, we vacate the court's temporary order granting <u>visitation</u> rights and remand for further proceedings.
Core Terms loco parentis, best interests of the child, grandparent, <u>custody</u>, <u>visitation</u> rights, nonparent, superior court, grandparent <u>visitation</u>, trial court, safeguards, best interest, fit parent, evidentiary, <u>visitation</u> order, reasonable <u>visitation</u> rights, great-grandparents, doctrine, constitutional right, third-party, temporary, convincing evidence, rebut a presumption, court's decision, liberty interest, same-sex, considerable discretion, step-grandparents, step-parent, non-parent, intrusion		BACKGROUND 1 P2 Michelle Egan and Therese Hochmuth were partners in a same-sex relationship for seventeen years. They agreed they wanted to have a child together and that Egan would give birth to the child. They made arrangements with a mutual friend to donate sperm. Egan was artificially inseminated, and the child was born in 2000. They jointly raised the child for seven years until their relationship ended in March 2007. At that point, they verbally agreed to alternate weeks with the child. In July, however, Egan expressed concern	
Case Summary Procedural Posture Petitioner biological mother sought review of, and a temporary stay of, an order of the Superior Court in Coconino County (Arizona) adopting a <u>visitation</u> plan that divided the child's time equally between the mother and respondent nonparent but also providing that she	LexisNexis® Headnotes Civil Procedure > Appeals > Standards of Review > De	Family Law > ... > <u>Visitation</u> Awards > Standards > Best Interests of Child Family Law > ... > <u>Visitation</u> Awards > Third Parties > Nonparents Family Law > ... > <u>Visitation</u> Awards > Third Parties > Nonparents Family Law > Child <u>Custody</u> > <u>Custody</u> Awards > General Overview Family Law > ... > <u>Visitation</u> > <u>Visitation</u> Awards > General Overview HN175-4: The legislature intended that courts treat <u>custody</u> and <u>visitation</u> petitions in a different manner, presumably because granting <u>visitation</u> is a far lesser intrusion, or assertion of control, than is an award of	¹For ease of reference, when we discuss a "person who <u>stands in loco parentis</u>" in the context of <u>§ 25-415</u>, we refer to such person as a "nonparent." See <u>A.R.S. § 25-415 ("Custody by nonparent")</u>; <i>infra</i> P 31. Similarly, we refer to a "legal [***2] parent," as defined in <u>§ 25-415(G)(2)</u>, as a "parent."

Figure 1: Sample Case Format of a U.S. State Appellate Court Judgment

The prompts were then designed to factor the document's structure. For example, when identifying cases involving individuals who were neither biological nor legal parents seeking custody, I instructed the model to read the case summary and, when unavailable, the opening paragraphs of the "Opinion." A qualitative review of more than 300 cases indicated that the nature of the dispute was almost always summarized in one of these locations.

By contrast, Indian judgments lacked these formal demarcations. Figure 2 presents a sample of how an India case transcript appears. Nonetheless, they informally followed a recognizable structure, typically opening with the background, presenting legal arguments, and closing with a decision. When the case length was within the context window, I directed the model to focus on the opening portions of the judgment to identify the nature of the dispute, and the closing portions to determine the outcome.

Dudela Radhakrishna Madhukarbhai Rao vs Manisha Radhakrishna Dudela D/O ... on 19 August, 2021 Dudela Radhakrishna Madhukarbhai Rao vs Manisha Radhakrishna Dudela D/O ... on 19 August, 2021 Equivalent citations: AIRONLINE 2021 GUJ 1550 Author: Ilesh J. Vora Bench: Ilesh J. Vora R/CR.RA/1299/2019 CAV JUDGMENT DATED: 19/08/2021 IN THE HIGH COURT OF GUJARAT AT AHMEDABAD R/CRIMINAL REVISION APPLICATION NO. 1299 of 2019 FOR APPROVAL AND SIGNATURE: HONOURABLE MR. JUSTICE ILESH J. VORA 1 Whether Reporters of Local Papers may be allowed to see the judgment ? 2 To be referred to the Reporter or not ? 3 Whether their Lordships wish to see the fair copy of the judgment ? 4 Whether this case involves a substantial question of Law as to the Interpretation of the Constitution of India or any order made thereunder ? DUDELA RADHAKRISHNA MADHUKARBHAI RAO Versus MANISHA RADHAKRISHNA DUDELA D/O VINODKUMAR KAR APPEARANCES: MR. D.B. REDDY, LD COUNSEL WITH MR. N.V. GANDHI (1693) for the Applicant MR. SUBRAMANIAM IYER (2104) for the Respondent No. 1 MRS KRINA CALLA APP for the Respondent(s) No. 2 Indian Kanoon - http://indiankanoon.org/doc/190782852/	Dudela Radhakrishna Madhukarbhai Rao vs Manisha Radhakrishna Dudela D/O ... on 19 August, 2021 (xii) It is denied that the petitioner is having monthly salary of Rs.8,000/- as on the day of filing the maintenance application, his salary was Rs.42,000/- p.m. In view of the aforesaid facts, the petitioner husband stated that the entire petition filed by the respondent wife is based R/CR.RA/1299/2019 CAV JUDGMENT DATED: 19/08/2021 on concocted lies and without any basis and therefore, may kindly be dismissed with cost. 3. Before the Family Court, Ahmedabad, the respondent wife adduced the evidence. She herself was examined at Exhs. 16 and 39. In support of her evidence, the witnesses namely Nilesh Kanubhai Maru at Exh.47, Pawan Bharathbhai Soni - Exh. 48, Priest Surjanarayan Murthy - Exh. 49, Father V.P. Kar and Mother Laxmben at Exh. 51 and 52, were examined. The testimony of the petitioner husband was recorded at Exh. 63 and in support thereof, he had examined Aslam Prabhakar - Exh. 70, Ajay Shankar at Exh. 74, Shri Hari Kumar Narsinh Rao - Exh. 75, Chanur Shaikh at Exh. 76, Kandula Srinivas Sulu at Exh. 77. 4. Parties have produced voluminous documents before the Family Court. The respondent wife has mainly relied on the photographs and video recording of the marriage, invitation card, profile of petitioner husband duly uploaded on the Jivasathi matrimonial site. 5. Upon considering the evidence and hearing the parties, the Family Court held that the marriage between the petitioner and respondent No.1 has R/CR.RA/1299/2019 CAV JUDGMENT DATED: 19/08/2021 taken place on 05.06.2014 as claimed by the wife. The learned Family Court recorded the findings of the fact that, the petitioner, in his profile, uploaded with the matrimonial site, has stated his status as divorcee and accordingly, the respondent wife relying the status of the husband agreed for marriage. The learned Family Court observed that, the respondent wife was unaware with regard to previous marriage with Toral Jhaveri at Jamnagar and same was suppressed by the petitioner husband. Thus, the learned Family Court relying on the decision of Badshah Vs. Urmila Badshah Godse, (2014) 1 SCC 188, held that if the wife is not aware of the previous marriage of the husband and she married to such husband, she is entitled to the maintenance. The learned Family Court has observed that strict proof of evidence is not required to prove the alleged marriage. The learned Family Court has taken into consideration the salary particulars of the petitioner husband and considering the fact that the respondent wife is unable to maintain herself, it is held that the petitioner husband having sufficient means, neglected respondent wife in providing maintenance and the petitioner husband ordered to pay monthly maintenance of Rs.40,000/- from the date of application. 6. Aggrieved by the impugned order of maintenance vide order dated 30.08.2019, the petitioner husband is R/CR.RA/1299/2019 CAV JUDGMENT DATED: 19/08/2021 before this Court by way of present Revision Application. 7. Heard learned counsel Mr. D.B. Reddy assisted by Mr. N.V. Gandhi, learned counsel for the petitioner, Mr. S.H. Iyer, the earned counsel for the respondent No. 1 wife and Mrs. Krina Calla, learned Ms. Krina Calla, ld. APP for the respondent No. 2 State. Indian Kanoon - http://indiankanoon.org/doc/190782852/	Dudela Radhakrishna Madhukarbhai Rao vs Manisha Radhakrishna Dudela D/O ... on 19 August, 2021 be read together harmoniously and on combined reading of both the sections, it becomes clear that, the marriage between two Hindus in contravention of Section 5(i) does not become void automatically. It becomes void only when it has declared void by a decree of nullity by the court of competent jurisdiction on a petition presented by the either party. The pre-conditions for a marriage being void under Section 11 are - (1) either party must make an application to the court of competent jurisdiction, (2) such application must be adjudicated by the court in accordance with the law and (3) the court must pass a decree declaring the marriage to be void. In absence of any such application made by either party and in absence of any adjudication on such application and in absence of decree of declaration, the marriage does not become void automatically. Reliance has been placed on the case A. Subash Babu Vs. State, reported in (2011) 7 SCC R/CR.RA/1299/2019 CAV JUDGMENT DATED: 19/08/2021 616, para 10 of the said judgment reads thus: "Though Section 11 of the Hindu Marriage Act provides that any marriage solemnized, if it contravenes the conditions specified in Clause (i) of Section 5 of the said Act, shall be null and void, it also provides that such marriage may on a petition presented by either party thereto, be so declared. Though the law specifically does not cast obligation on either party to seek declaration of nullity of marriage and it may be open to the parties even without recourse to the Court to treat the marriage as a nullity, such a course is neither prudent nor intended and a declaration in terms of Section 11 of the Hindu Marriage Act will have to be asked for, for the purpose Reportable of precaution and/ or record. Therefore, until the declaration contemplated by Section 11 of the Hindu Marriage Act is made by a competent Court, the woman with whom second marriage is solemnized continues to be the wife within the meaning of Section 494 IPC and would be entitled to maintain a complaint against her husband." In view of the aforesaid, it is the submission of learned counsel that, in the present case, neither petitioner nor respondent had moved any petition before the competent court for obtaining decree of declaration as contemplated by Section 11 of the Hindu Marriage Act, 1955 and therefore, the marriage between the parties cannot be held invalid and the respondent wife is legally entitled for the maintenance. 18. Learned counsel for the respondent wife has further submitted that, strict proof about valid marriage is not sine qua non for getting a maintenance under Section 125 of the Code. The oral evidence of the respondent wife and supporting testimonies of the witnesses as R/CR.RA/1299/2019 CAV JUDGMENT DATED: 19/08/2021 well the documentary evidence like photographs, video recording of marriage, invitation card would demonstrate that, the marriage has taken place on 05.06.2014 and thereafter, the parties lived together and cohabitation has taken place. Under the circumstances, the presumption arises in favour of the wife that marriage has taken place for which the petitioner failed to rebut the said presumption. Thus, the learned Family Court after analyzing the evidence adduced by the parties has rightly decided the issues and held the contention negatived raised by the petitioner husband for which no interference is required. Indian Kanoon - http://indiankanoon.org/doc/190782852/
--	---	---

Figure 2: Sample Case Format of an Indian High Court Judgment

When judgments exceeded the context window, a different strategy was necessary. Researchers often address this problem by chunking texts into smaller segments (J. H. Choi 2024). The drawback is that relevant information may be distributed across multiple chunks. To mitigate this risk, I generated multiple strategically selected samples of the text based on the particular task. For outcome classification, for example, I provided several overlapping excerpts from the concluding portions of a case - including the final 20 percent of the text, the last 3,000 characters, and the final ten paragraphs - and asked the model to classify the outcome and report its confidence for each excerpt.

The broader point is that document format can be leveraged to ensure that the model focuses on the “right” information. Whether that structure is formal, as in U.S. appellate opinions, or informal, as in Indian judgments, understanding where the relevant information is likely to appear allows researchers to provide LLMs with more helpful context and obtain better-grounded classifications.

3.1.2. DESIGNING PROMPTS GIVEN THE TYPE OF INFORMATION REQUIRED

To determine whether the information is relevant or not, researchers must also consider what type of information they want the LLM to collect and classify. In my experience, court data projects involve two distinct tasks: extracting explicitly stated information and inferring more complex concepts from information distributed throughout a judgment. Because these tasks place different “cognitive” demands on the model, they demand different prompting strategies.

Extraction tasks involve collecting information that is explicitly stated in the judgment, such as litigant gender, cohabitation, or caregiving activities. Because these details are typically described directly, multiple variables can often be collected through a single prompt. In the U.S. dataset, for example, one prompt extracted information on caregiving duration, claimant privilege, and criminal history. The full prompt is inserted below:

Extraction Prompt: Caregiving Extraction

You are an American legal analyst with a specialization in family law.
Extract evidence of the third-party claimant’s caregiving role and
relationship with the child.

CONTEXT FROM STAGE 1:

- Third-Party Claimant: {third_party_claimant}
- Opposition: {opposition}

Read the complete case and extract:

=====

THIRD-PARTY CLAIMANT <-> CHILD RELATIONSHIP

=====

- Caregiving Duration: For how long did the claimant provide care to
the child? (<6 months / 6-24 months / 2-5 years / 5+ years / Unknown)
- Caregiving Duration Evidence: Direct quote or "Not found"
- Cohabitation: Did the claimant live with the child?

(0 = No co-residence / 1 = Partial or intermittent /
2 = Full-time co-residence)

- Residential Caregiving Evidence: Direct quote describing living arrangements or household integration, or "Not found"
- Caregiving Depth: What caregiving functions did the claimant perform?
(0 = Limited support only / 1 = Daily care without major decision-making /
2 = Daily care plus school/medical/disciplinary decisions)
- Caregiving Depth Evidence: Quote describing feeding, bathing, supervision, schooling, medical care, discipline, or other parental functions, or "Not found"

=====

THIRD-PARTY CLAIMANT'S PRIVILEGE & SOCIAL CONTEXT

=====

- Claimant Privilege Indicators: Evidence of education, disability, race, class, homelessness, socioeconomic, or immigration status
(High/Medium/Low/Unknown)
- Claimant Privilege Evidence: Direct quote, or "Not found"
- Claimant Criminal History: Has claimant served time/been in prison?
(Yes/No/Unknown)
- Claimant Criminal History Evidence: Direct quote, or "Not found"

By contrast, inference tasks require the model to synthesize information distributed throughout a judgment and classify a more complex concept. Examples include determining whether a case involved a woman seeking recognition as a wife despite the existence of a prior marriage, whether a claimant demanded child custody despite lacking legal status, and whether a judgment ultimately benefited a particular individual. Because these classifications depend on multiple facts, legal arguments, and judicial conclusions, I found it more efficient to dedicate an entire prompt to

a single inferential task. A prompt for such a task is shown below:

Inference Prompt: Indian High Court Outcome Determination

You are a senior Indian legal expert analyzing an Indian High Court judgment. Your task is to determine the EXACT outcome of this case.

Based ONLY on the text above, determine the court's decision.

Pay special attention to:

- The FINAL paragraph which typically contains the court's order.
- Phrases like "petition is allowed/dismissed", "appeal is allowed/dismissed", "application is allowed/dismissed", "not maintainable", "hence it is dismissed", "accordingly allowed", "writ petition is allowed/dismissed", etc.
- The concluding sentence, often starting with phrases like "In view of", "Accordingly", "Therefore", etc.
- If there are different outcomes for different petitioners, note this as a split decision.
- Phrases that indicate the disposition such as "is entitled to", "be released on", "is granted", "is rejected", "is disposed of".

Tailoring prompts based on the type of information that was being collected improved the clarity of the task presented to the model. It reduced ambiguity and made it easier to evaluate whether the model was working with the relevant information, particularly in cases that required substantial inference.

3.2. ENSURING RELIABLE OUTPUTS

Effective use of LLMs requires not only feeding the model relevant information but also minimizing hallucinations and biases. While recent research suggests that LLMs can classify legal texts at levels comparable to human coders, they remain capable of generating inaccurate information.

A common safeguard is to reduce the model’s temperature. When accessed through APIs, temperature typically ranges from 0 to 2, with lower values limiting the scope of its interpretation and higher values encouraging more creative and exploratory responses. In studies that attempt to extract and classify information, the temperature is typically set to 0.

Besides lowering the temperature, I relied on three additional practices: requiring textual evidence for every classification, validating outputs through multiple classification strategies, and adopting conservative coding rules when information was ambiguous.

3.2.1. REQUIRING TEXTUAL EVIDENCE FOR CLASSIFICATIONS

For every classification, the model was required to provide supporting textual evidence. For example, when collecting information on the circumstances surrounding which an individual claimed child custody, the model not only classified the trigger (e.g., state intervention or parent cutting contact) but also extracted the passage supporting that conclusion. Figure 3 provides a sample of these outputs. This made classifications transparent and allowed for quick verification without reading the entire opinion.

Trigger_Event	Trigger_Event_Evidence
Parent's Imprisonment	Danielle stated that Tamara had been incarcerated and her current address was unknown.
State Intervention	"The Department became involved with Mother and the children again in May 2021."
Parent Cuts Contact	"When the child was six years old, the mother and the co-habitant discontinued their relationship. Thereafter, the cohabitant instituted a custody proceeding."
State Intervention	"The child was again taken into protective custody in July 2008 after she was found alone in a busy intersection with a soiled diaper."
State Intervention	"As a result of that arrest and Mother's admission that she had a drug problem and was unable to provide a stable home for the baby, the Cabinet filed a juvenile dependency, neglect and abuse petition on March 9, 2007, in Lewis Family Court."
Parent Cuts Contact	Appellant testified that Joseph "stopped allowing me to see [S.F.]" (Tr. at 13.)
Parent Cuts Contact	"However, twenty months later, appellant filed for divorce on August 17, 2010. In his answer and counterclaim for divorce, appellee requested visitation with S.B., alleging that he was entitled to this privilege because he had stood in loco parentis to the child for the past five years."
Parent Cuts Contact	"When Dr. Fox was permitted visitation pending the hearing, he saw both S.G. and J.F. Ms. Glassing required that any visitation be by court order."
Parent Cuts Contact	"Two months after the agreement was signed, Thomas filed a contempt petition stating that he had not seen the twins in two months and alleging that Caldwell had failed to make reasonable efforts to facilitate his relationship with the twins."
Parent Cuts Contact	"Mother also prevented Grandmother from having contact with the children."
State Intervention	"the initiation of this case by DHS, when the children were removed from their mother due to her drug use."
State Intervention	"Petitioners filed for temporary custody of the child in June 2014. The court granted their request in August 2014, and ordered Mother to start participating in random drug testing through TASC immediately."
Parent Cuts Contact	"In 2001, after the parties' relationship ended and a dispute over parenting time arose, Clark sought to restrict McLeod to ten overnights per month in 2003 and six overnights per month in 2004, and to terminate all court-ordered parenting time in 2005."
Parent's Health Issue	"On February 18, 2009, she checked herself into a mental health facility. But before doing so, she arranged ... to have Patti pick up the children."
State Intervention	"In September 2016, the Santa Cruz County Human Services Department (Department) filed a petition under Welfare and Institutions Code section 300... relative to an infant boy, D.L. (the minor). ... The Department removed the minor from the custody of mother and father ... and he was placed in a foster home with D.C. and B.C."
State Intervention	"In December 2009, the juvenile court asserted dependency jurisdiction over then one-year-old J.T. under section 300, subdivision (b). ... The court found mother left J.T. with his maternal grandmother without making appropriate plans for his care, and mother's whereabouts were unknown."
State Intervention	"On December 14, 2005, the Department of Children and Family Services (DCFS) received a referral regarding then three-year-old Joshua and two-year-old Nicholas due to domestic violence between Father and Mother..."
State Intervention	"On September 22, 2022, the department filed a dependency petition ... alleging ... father physically abused maternal half sibling A.X.... [and] the parents exposed the children to domestic violence."
State Intervention	"The juvenile court granted DCFS authorization to remove Richard on July 12, 2013, and the social worker detained Richard on July 15, 2013 from mother"
State Intervention	"Mother's and father's whereabouts were unknown when Brandon, then age 2, his infant sister Taylor, 7 months, and their half-brother Michael, 7 years, were first taken into protective custody and placed in an emergency shelter home in February 1999."
Parent's Health Issue	"Then nine-year-old A.L. was taken into protective custody by Orange County Social Services Agency (SSA) in June 2009, due to neglect and caretaker absence arising from Mother's mental illness. Mother had been involuntarily hospitalized under section 5150..."

Figure 3: Example of an Auditable Classification

3.2.2. VALIDATING OUTPUTS THROUGH MULTIPLE CLASSIFICATION STRATEGIES

To avoid further classification errors, I incorporated another practice which involved classifying the same concept using multiple approaches and comparing the results. This practice was especially helpful when collecting information on case outcomes. I used two distinct approaches to classify this variable. First, I asked the model to assign the case to a predetermined list of outcomes such as “Custody Granted,” “Spousal Support Granted,” or “Visitation Denied.” Second, I asked the model to provide an open-ended response on whether the individual was broadly “Recognized,” “Not Recognized” or “Unclear” by the court. Cases in which these two outputs diverged were flagged for manual review. This approach proved particularly useful in cases where the disposition was ambiguous or involved multiple forms of relief.

3.2.3. CONSERVATIVE CODING APPROACH

Case judgments are interpretive documents rather than comprehensive repositories of facts. Judges routinely rely on information that never appears explicitly in court opinions. For example, in the India project, judges could likely infer the caste background of many litigants but caste was explicitly mentioned in only 112 of 2,721 cases in the dataset. Asking the model to infer caste in the remaining cases would therefore require relying on assumptions rather than textual evidence. Unless these inferences can be validated using independent data, relying solely on LLM to generate them is difficult to justify, particularly given the possibility of hallucination.

For this reason, the model was instructed to code a feature only when it was explicitly supported by the judgment. When information was ambiguous or absent, the default classification was “Unknown” rather than an inferred value. While this approach likely underestimates the prevalence of certain characteristics, it reduces the risk of hallucination and ensures the classifications remain grounded in the text being analyzed.

3.3. ACCURACY AGAINST HUMAN CODING

To assess whether these practices produced accurate classifications, I hand-coded a random sample of 300 cases drawn from each dataset and compared the hand-coded classifications to the model’s output. Overall, the model’s classifications matched human coding in 95% of the classifications.

When the information extracted was directly stated in the text such as litigant gender or whether there was a marital ritual, the agreement was 97%. The cases where there were discrepancies were primarily because there was disagreement between the parties on case details. For example, in a case where a woman claimed that there had been a marital ritual, the opposing party argued that there had been no such ritual and there was no evidence to support the claim. Wherever possible, in such cases, I noted that the value should be “Unknown.” By contrast, when the information extracted depended on inference, i.e. classification based on assembling multiple facts distributed across the judgment, the agreement rate was 92%. Inference-based extraction was mainly for determining whether the case was relevant and identifying the case outcome. The accuracy rate with respect to case relevance was 94%, which is notably high: Of 600 cases, only 36 were incorrectly selected. In comparison, the accuracy rate in relation to case outcome was relatively lower at 90%. These cases had been flagged by the two-step check described above.

Reviewing these discrepancies revealed that the issues were less due to the model and more due to ambiguous dispositions and required close, multiple readings of the case document to determine the outcome.

In sum, these results suggest that the framework produces classifications accurate enough for large-scale descriptive analysis, further suggesting that with this framework, LLMs can be a powerful tool to analyze case data at scale while also ensuring that their outputs are verifiable.

4. ETHICAL CONSIDERATIONS

Despite their advantages over traditional pattern matching, LLMs raise distinct ethical concerns. As Bender et al. note (2021), “the size of the training data they leverage and the spheres of influence they can possibly affect” means that it “incurs new kinds of risks.” Among the most important are environmental costs, loss of opportunities for student RAs, and the further concentration of power in a small number of firms that control the development and deployment of LLMs. Using LLMs over standard pattern matching approaches should be considered in light of these risks as well.

First, training and operating LLMs requires substantial energy consumption (Strubell, Ganesh, and McCallum 2019). As the journalist Stephen Witt (2025) notes, “To write a college term paper, an A.I. might produce about five thousand tokens, consuming enough electricity to run a microwave over at full power for about three minutes.” These costs extend beyond electricity use to the construction and operation of data centers, which can impose environmental burdens on surrounding communities (Hassan 2026). Researchers should therefore weigh these costs against the expected benefits of using LLMs. For relatively straightforward tasks, such as identifying litigant gender or case subject, traditional pattern matching may remain preferable. By contrast, researchers interested in concepts that are more deeply embedded in the text may benefit from the framework outlined above. Such an approach that prioritizes feeding LLMs the “right” information makes it possible to leverage the strengths of LLMs while also ensuring that they are not “endlessly data hungry” (Bender et al. 2021).

Second, as Lee et al. (2025) note, there is a risk that the use of LLMs, as presented in this article, will “outsource tasks traditionally performed by student RAs.” At the same time, validation continues to require substantial human oversight. Moreover, by lowering technical barriers to large-scale text analysis, LLMs make computational research more accessible to scholars whose

expertise lies outside programming and data science, enabling them to participate in conversations that have historically been dominated by researchers with specialized computational skills.

Third, using off-the-shelf LLMs including OpenAI's GPT-4 and Claude's Sonnet 4.5 risks centralizing "[a] handful of companies' control" over a "\$15.7 trillion game changer of an industry" (Weil 2023). For researchers, this raises concerns with respect to safety and privacy. There is no clarity on what happens to data that is fed into the system and equally, there is no accountability in terms of what data the model is relying on to process and classify information (Spirling 2023). This means that researchers should be cautious in using LLMs particularly when case data is not public or contains sensitive information.

Outside of concerns over safety and privacy, there is a deeper public fear that AI is an "elite political project" that is being "pushed down our throats" (The Ezra Klein Show, New York Times Podcasts, and New York Times Opinion 2026). Arguably, using LLMs then can ostensibly be a political statement. However, an outright rejection of LLMs is both difficult to sustain and arguably, counterproductive given their widespread adoption by consumers and businesses. In fact, many policymakers today are focused on developing regulatory frameworks informed by a nuanced understanding of how these systems operate in practice.⁶

In this context, the approach outlined in this article offers one way for researchers to engage with LLMs in a manner that allows them to develop a more textured understanding of both their capabilities and limitations, making them better equipped to evaluate the promises and shortcomings of AI.

⁶Alex Bores, a New York State assembly member, is renowned for his interest in regulating AI, and not banning it, to an extent that companies like OpenAI, Google, and Palantir have targeted him with ads that assault his reputation. Bores' regulatory recommendations stem from his own personal experience working at these companies and his understanding of how AI works: <http://nytimes.com/2026/06/06/nyregion/alex-bores-primary-ai-palantir.html>

REFERENCES

- Ash, Elliott et al. (2025). “In-Group Bias in the Indian Judiciary: Evidence from 5 Million Criminal Cases”. *The Review of Economics and Statistics*, pp. 1–45.
- Bender, Emily M. et al. (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. New York, NY, USA: Association for Computing Machinery, pp. 610–623.
- Bhupatiraju, Sandeep et al. (2024). “Legal Disparities and Judicial Mitigation: Muslim Petitioners in an Indian High Court”.
- Choi, Donghyun Danny, J. Andrew Harris, and Fiona Shen-Bayh (2022). “Ethnic Bias in Judicial Decision Making: Evidence from Criminal Appeals in Kenya”. *American Political Science Review* 116.3, pp. 1067–1080.
- Choi, Jonathan H. (2024). “How to Use Large Language Models for Empirical Legal Research”. *Journal of Institutional and Theoretical Economics* 180.2, pp. 214–233.
- Deroy, Aniket, Kripabandhu Ghosh, and Saptarshi Ghosh (2024). *Applicability of Large Language Models and Generative Models for Legal Case Judgement Summarization*.
- Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). “ChatGPT outperforms crowd workers for text-annotation tasks”. *Proceedings of the National Academy of Sciences of the United States of America* 120.30, e2305016120.
- Grimmer, Justin and Brandon M. Stewart (2013). “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. *Political Analysis* 21.3, pp. 267–297.
- Hassan, Adeel (2026). “The Cloud Has Sound: The Unrelenting and Unseen Cost of A.I. Data Centers”. *The New York Times*.
- Heddaya, Mourad et al. (2024). *CaseSumm: A Large-Scale Dataset for Long-Context Summarization from U.S. Supreme Court Opinions*.
- Joshi, Shareen et al. (2024). *Caste Aside? Names, Networks and Justice in the Courts of Bihar, India*. SSRN Scholarly Paper. Rochester, NY.
- Lazarev, Egor (2019). “Laws in Conflict: Legacies of War, Gender, and Legal Pluralism in Chechnya”. *World Politics* 71.04, pp. 667–709.

- Lee, Kyuwon et al. (2025). “Applications of GPT in Political Science Research: Extracting Information from Unstructured Text”. *PS: Political Science & Politics* 58.4, pp. 630–640.
- Liebman, Benjamin L. et al. (2017). “Mass Digitization of Chinese Court Decisions: How to Use Text as Data in the Field of Chinese Law”. *SSRN Electronic Journal*.
- Lucas, Christopher et al. (2015). “Computer-Assisted Text Analysis for Comparative Politics”. *Political Analysis* 23.2, pp. 254–277.
- Pavlus, John (2024). *Does AI Actually Understand Language?*
- Somers, James (2025). “The Case That A.I. Is Thinking”. *The New Yorker*.
- Spirling, Arthur (2023). “Why open-source generative AI models are an ethical way forward for science”. *Nature* 616.7957, pp. 413–413.
- Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). “Energy and Policy Considerations for Deep Learning in NLP”. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3645–3650.
- The Ezra Klein Show, New York Times Podcasts, and New York Times Opinion (2026). *Why Are Palantir and OpenAI Scared of Alex Bores? | The Ezra Klein Show*.
- Truscott, Jake S. and Michael K. Romano (2026). “Measuring Judicial Ideology Through Text”. *Journal of Law and Courts* 14.1, pp. 65–82.
- Weidmann, Nils B., Mats Faulborn, and David García (2026). “Large Language Models Are Democracy Coders with Attitudes”. *PS: Political Science & Politics* 59.1, pp. 17–23.
- Weil, Elizabeth (2023). “You Are Not a Parrot”. *Intelligencer*.
- Witt, Stephen (2025). “Inside the Data Centers That Train A.I. and Drain the Electrical Grid”. *The New Yorker*.