

FIRAC-APPEALS: A Structured Indian Appellate Judgments Dataset for Summarisation and Reasoning-Aware Legal NLP

Martin Borčín Joemon M. Jose

University of Glasgow

Glasgow, United Kingdom

{2481234b, joemon.jose}@glasgow.ac.uk

Abstract

Legal natural language processing (NLP) resources often flatten judgments into full texts paired with summaries or outcome labels, despite appellate reasoning being organised around facts, issues, authorities, analysis, and dispositions. We introduce FIRAC-APPEALS, a public dataset and annotation pipeline that converts 45,020 Indian Supreme Court appeals into machine-generated, silver-standard FIRAC case briefs with metadata, best-effort leakage-reduced Facts, typed Rules/Authorities, Analysis, Conclusion, and outcome labels. This structure supports headnote generation, issue spotting, authority recommendation, reasoning-aware audit, and leakage-aware outcome prediction. Human legal assessment, independent-model screening, summarisation comparisons, and downstream baselines provide evidence of the resource’s usability beyond conventional summarisation or outcome-only legal judgment prediction, while also exposing residual authority, reference-dependence, and leakage risks.

1 Introduction

Appellate judgments distribute legally relevant material across facts, procedure, issues, authorities, reasoning, and disposition. Yet many legal NLP resources expose only full judgments paired with a summary or outcome label, making it difficult to reuse public cases for issue spotting, authority grounding, knowledge-graph construction, leakage-aware prediction, and diagnosis of reasoning failures.

This is especially problematic for prospective legal judgment prediction (LJP) and reasoning tasks. A completed judgment is written after the outcome is known, so the full text contains pervasive leakage: holdings, justificatory reasoning, procedural posture, and dispositive tone all encode the court’s view. Masking the final label therefore cannot remove the broader outcome signal, making conclu-

sions from full-judgment LJP benchmarks difficult to interpret, as emphasised in critiques of LJP evaluation (Medvedeva et al., 2023; Medvedeva and McBride, 2023). Sentence-level rhetorical-role resources add useful structure (Nigam et al., 2025), but extractive spans remain authored as part of the court’s final judgment and can still carry stance-bearing language.

We introduce FIRAC-APPEALS, a structured resource and annotation pipeline for Indian appellate judgments. The central idea uses FIRAC as a lightweight semantic interface between raw judgments and downstream legal NLP tasks. Facts provide neutral pre-decision inputs. Issues, Rules/Authorities, Analysis, and Conclusion provide reusable machine-generated references for summarisation, retrieval, reasoning, and outcome prediction.

Rather than forcing all tasks through full-text inputs and final labels, FIRAC-APPEALS exposes intermediate components that can help diagnose which part of a system succeeds or fails, while also enabling the construction of diverse tasks for evaluating or adapting systems across different legal NLP settings.

We contribute: (i) FIRAC-based structuring as a lightweight resource-design principle for legal NLP; (ii) a 45,020-case Indian appellate resource with machine-generated FIRAC annotations, outcomes, authorities, metadata, and scripts; and (iii) assessment through corpus statistics, quality and leakage audits, summarisation comparisons, and baselines for issue spotting, authority recommendation, and outcome prediction.

2 Background

Prior legal NLP resources usually optimise for one task at a time. Legal judgment prediction datasets such as ILDC (Malik et al., 2021) pair case materials with outcomes and, in some settings, ex-

Table 1: Positioning of FIRAC-APPEALS against related legal NLP resources. The comparison concerns resource affordances rather than benchmark scores. “Structured legal fields” means annotations beyond a single unstructured summary or final outcome label.

Resource	Jurisdiction court	/ Size	Primary target	Structure	Outcome	Authority signal	Leakage-aware input
ILDC (Malik et al., 2021)	India / SC	34,816	Prediction + explanation	56 expert rationales	Yes	None / implicit	Partial
IN-Abs (Shukla et al., 2022)	India / SC	7,130	Abstractive headnotes	None	No	None / implicit	No
IN-Ext (Shukla et al., 2022)	India / SC	50	Extractive summaries	Rhetorical roles	No	Statute / precedent labels	No
UK-Abs (Shukla et al., 2022)	UK / SC	793	Press summaries	3 press sections	No	None / implicit	No
LawSum (Parikh et al., 2021)	India / SC	10,764	Headnotes	Metadata; weak labels	No	Citations + acts	No
CivilSum (Malik et al., 2024)	India / SC + HCs	23,350	Civil summaries	Paragraph references	No	References + legislation	No
MILDSum (Datta et al., 2023)	India / SC + HCs	3,122	English-Hindi summaries	Bilingual pairs	No	None / implicit	No
FIRAC-APPEALS	India / SC	45,020	FIRAC headnotes + tasks	F/I/R/A/C	Yes	Typed authorities	Best-effort

planations, while summarisation resources such as IN-Abs, IN-Ext, and UK-Abs (Shukla et al., 2022), LawSum (Parikh et al., 2021), CivilSum (Malik et al., 2024), and MILDSum (Datta et al., 2023) provide judgment-summary pairs, metadata, paragraph references, or multilingual summaries. As Table 1 shows, these resources are valuable but leave a gap between outcome labels and unstructured summaries: few expose reusable issue, authority, reasoning, and leakage-aware input fields. This matters because legal annotation is itself interpretive: disagreement between trained annotators can be informative, and dataset papers should report the annotation process and agreement structure rather than silently collapsing it into a single gold view (Braun, 2024).

Recent legal NLP increasingly evaluates models on reasoning rather than only text matching or final labels. Statutory-reasoning, LegalBench, and LawBench tasks test rule application, legal entailment, issue classification, and multi-step inference (Blair-Stanek et al., 2023; Guha et al., 2023; Fei et al., 2024). Critiques of legal judgment prediction (LJP) stress realistic inputs, leakage control, and evaluation beyond headline accuracy (Medvedeva et al., 2023; Medvedeva and McBride, 2023). Issue, Rule, Application, Conclusion (IRAC) and FIRAC provide compact legal-practice scaffolds for this missing middle (Sinclair, 2003; Turner, 2015): related work uses IRAC-style structure for issue identification, rule retrieval, application, and conclusion generation (Kang et al., 2025), graph-constrained Indian legal reasoning systems use IRAC nodes to verify citation paths and detect unsupported precedents (Bose, 2026), and formal or neuro-symbolic LJP work similarly seeks explicit rule-application structure (Wei et al., 2025; Zhang et al., 2025). FIRAC-APPEALS complements these lines by applying FIRAC to real appellate judgments at scale. Figure 1 also shows that the Supreme Court subset

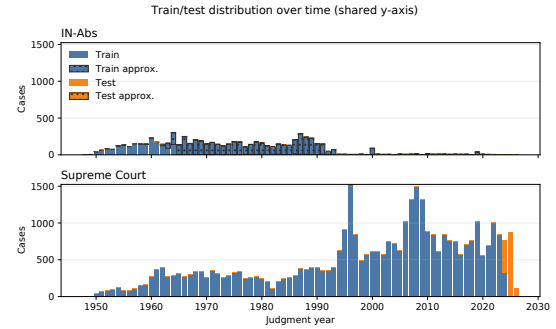


Figure 1: Temporal distribution of judgments. The Supreme Court subset is larger, more recent, and supports temporal evaluation. IN-Abs years are approximate when dates are missing.

expands the size, recency, and temporal evaluation range of available Indian appellate data.

3 Dataset Creation and Contents

FIRAC-APPEALS is built from two public upstream sources: IN-Abs (Shukla et al., 2022), distributed as judgment texts with reference headnotes, and the regularly updated Indian Supreme Court judgments repository vanga/indian-supreme-court-judgments. The upstream IN-Abs archive contains 7,130 judgment-summary pairs, split into 7,030 train and 100 test cases (four train cases failed annotation validation, so the processed FIRAC-APPEALS subset contains 7,126 IN-Abs rows). For the Supreme Court source, judgment text was extracted from source PDFs/records, deduplicated, normalized, and aligned with available source metadata. The resulting pipeline can be rerun as upstream Supreme Court cases are updated.

Figure 2 gives the high-level annotation flow. Each judgment was processed by specialised LLM roles: an issue analyst identifies dispositive questions, a legal researcher extracts statutes, precedents, and legal principles, a logician structures the

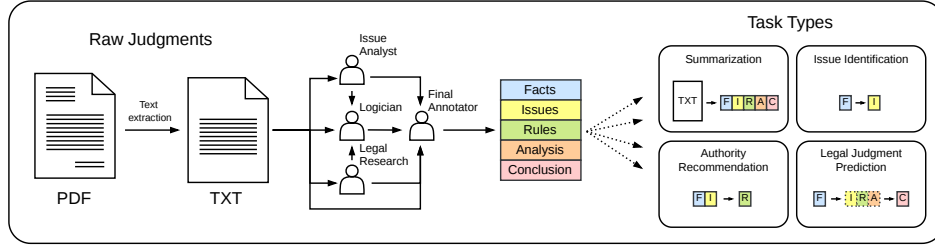


Figure 2: The pipeline. Raw appeal judgments are turned into structured artefacts, enabling their use for other legal NLP tasks.

court’s rule-to-fact application, and a final FIRAC judge reconciles these outputs into one JSON annotation. Facts were prompted as neutral pre-decision material, excluding holdings, reasoning, and outcomes. A separate metadata extractor records case identifiers, court, dates, parties, lawyers, and bench information where available. All roles used Qwen3.6-27B with thinking disabled¹. Consequently, role separation does not provide independent annotations. The model was selected as an open-weight system that could be served locally at corpus scale and reliably used with the pipeline’s structured outputs. We did not run a controlled pilot comparing alternative annotation generators, so model choice remains a dataset-design limitation. The simplified outcome label is emitted with the model-generated Conclusion and normalized to allowed, dismissed, or other. The resulting FIRAC records are machine-generated annotations rather than legal ground truth.

The dataset is publicly available on Hugging Face,² and the code, prompts, sanitized evaluation outputs, and reproducibility artifacts are released on GitHub.³ The dataset contains `in_abs`, `supreme_court`, and combined configurations. Each row contains stable identifiers, source provenance, split membership, judgment text, metadata, a nested FIRAC object, and a simplified outcome label. IN-Abs rows also include the original reference summary. Appendix A describes the schema in more detail, and Table 2 summarises the corpus.

The Supreme Court split is temporal: training decisions precede 1 July 2024, while test decisions are on or after that date (3 July 2024–4 May 2026 in the released data). Before binary filtering, train outcomes are 56.1% allowed, 39.1% dismissed,

Table 2: Dataset overview.

Statistic	IN-Abs	Supreme Court (SC)
Cases	7,126*	37,894
Train / test	7,026 / 100	36,453 / 1,441
Years	1948–2022	1950–2026
Median judgment words	3,147	3,248
Median FIRAC words	992	964
Compression ratio	0.36	0.36
Novel 2/3-grams	43.2 / 66.9%	44.1 / 68.1%
Outcome A/D/O	47.5 / 50.4 / 2.1%	56.6 / 38.5 / 4.9%
Reference summaries	100.0%	0.0%
Date coverage	99.5% (28.1% exact)	100.0%
Rules/case P/S/Pr	1.7 / 2.8 / 1.2	1.9 / 2.5 / 1.1

* Four train cases (of original 7,130) failed annotation.
A/D/O denotes allowed/dismissed/other.
P/S/Pr denotes precedents/statutes/principles.

and 4.8% other, compared with 70.4%, 23.7%, and 5.9% in test. Thus, evaluation combines temporal generalization with a substantial label shift and potentially with changes in legal topics and drafting style. Four Supreme Court rows remain in the 45,020-row release but lack complete posthoc records. Posthoc-dependent audits and tasks therefore use 45,016 rows where noted.

4 Dataset Quality Evaluation

4.1 Human Evaluation Design and Rubric

We evaluate annotation quality with three human legal annotators on 120 distinct cases from the held-out test subsets. The annotators are trained legal professionals with formal legal education who currently practise in a legal-technology setting. A1, A2, and A3 have approximately 1, 20, and 5 years of experience, respectively. The sample was stratified by subset and outcome, with 30 IN-Abs allowed, 30 IN-Abs dismissed, 30 Supreme Court allowed, and 30 Supreme Court dismissed cases. We collected 150 annotations in total: 90 cases received one annotation and 30 cases received two independent annotations. The paired overlap was balanced across annotators, with 10 shared cases for each annotator pair. A separate set of 10 training cases was provided for practice before the evaluation. Ratings were completed independently and

¹<https://huggingface.co/Qwen/Qwen3.6-27B>

²<https://huggingface.co/datasets/mborcin/firac-appeals>

³<https://github.com/MartinBorcin/firac-appeals-repro>

Table 3: Human rubric-score distribution. Percentages are case-weighted so each of the 120 cases contributes equally, with 150 annotations collected.

Component	Dimension	Ratings (%)						Mean [95% CI]
		1	2	3	4	5	< 4	
Facts	Correctness	0.0	1.7	0.0	6.2	92.1	1.7	4.89 [4.80, 4.96]
	Coverage	0.0	0.4	1.7	16.7	81.2	2.1	4.79 [4.70, 4.86]
Issues	Correctness	0.0	0.0	0.8	4.6	94.6	0.8	4.94 [4.88, 4.98]
	Coverage	0.0	0.0	1.7	13.3	85.0	1.7	4.83 [4.76, 4.90]
Rules/Authorities	Correctness	0.0	1.2	6.2	10.8	81.7	7.5	4.73 [4.61, 4.83]
	Coverage	0.0	2.1	7.1	23.3	67.5	9.2	4.56 [4.44, 4.68]
Analysis	Correctness	1.2	0.4	1.7	5.0	91.7	3.3	4.85 [4.74, 4.94]
	Coverage	0.4	1.2	2.1	18.8	77.5	3.8	4.72 [4.61, 4.81]
Conclusion	Correctness	0.8	0.8	0.0	6.2	92.1	1.7	4.88 [4.78, 4.96]
	Coverage	0.8	0.0	10.0	15.0	74.2	10.8	4.62 [4.49, 4.74]
Overall		0.3	0.8	3.1	12.0	83.8	4.2	4.78 [4.72, 4.84]

were not adjudicated. No formal post-practice calibration statistic was recorded.

Annotators evaluated the generated FIRAC structure against the reference judgment using component-specific guidelines released with the code. The rubric scored the five FIRAC components (Facts, Issues, Rules/Authorities, Analysis, and Conclusion) on two 1–5 dimensions: faithfulness/correctness, meaning grounding in the judgment and absence of unsupported, misleading, outcome-leaking, or wrongly framed content; and coverage/usefulness, meaning completeness for downstream legal reasoning, summarisation, and audit. Scores 1, 2, and 3 correspond to unusable, serious, and noticeable problems, while 4 indicates minor issues and 5 indicates a faithful, complete, and appropriately scoped section. Any score below 4 required a concise explanation. The rubric explicitly treats Rules/Authorities broadly, covering statutes, constitutional provisions, regulations, precedents, doctrines, tests, standards, and other legally operative principles considered by the court. It also specifies that partly allowed outcomes are mapped to the coarse label allowed, with detailed relief, remand, costs, and qualifications assessed through the Conclusion text. For Facts, annotators considered outcome leakage as one possible correctness defect alongside unsupported, misleading, or wrongly framed content.

Primary quality estimates are case-weighted: for double-annotated cases, we average annotators within case and then average cases, so the overlap subset does not receive twice the weight. Confidence intervals use case-bootstrap resampling. Inter-annotator agreement is computed only on the 30 double-annotated cases and reported with exact

agreement, within-one agreement, mean absolute error, Spearman correlation, ordinal Krippendorff’s α , and low-score detection overlap.

4.2 Human Evaluation Results

Table 3 shows high human-rated quality across the stratified sample. The case-weighted overall mean is 4.78, with 83.8% of ratings at 5 and only 4.2% below 4. Facts, Issues, Analysis, and Conclusion correctness are especially strong, all at or above 4.85 on average. Rules/Authorities has the lowest mean scores, especially for coverage (4.56; 9.2% below 4), while Conclusion coverage has the highest human below-4 rate (10.8%). The latter mostly reflects cases where the broad outcome was correct but some relief details, remand directions, costs, or qualifications were compressed.

The accompanying explanations for scores below 4 describe the character of the few lower-scoring cases rather than the typical annotation: they account for 66 score-level comments out of 1,500 human rubric scores. Annotators’ overall debrief comments described the outputs as mostly accurate, with occasional minor omissions and a small number of hallucination concerns. Among the lower-scoring annotations, Rules/Authorities comments mainly involved incorrect, invented, or source-unverifiable citations and omitted material precedents, precedent chains, doctrines, or minority opinions. Conclusion comments mostly concerned missing costs, remand directions, deadlines, ancillary orders, qualifications, or partial dispositions. Analysis comments involved omitted issue-level reasoning, rebuttals, independent grounds, and occasional confusion in split-opinion cases. Facts and Issues attracted fewer comments and were usually

affected by missing procedural or party context rather than affirmative errors. Thus, the qualitative feedback is best read as a description of residual failure modes in a small tail of difficult cases, especially authority provenance and split opinions. Because citation provenance can matter disproportionately in authority-sensitive downstream uses, Section 4.3 describes a patch that removes authority references that cannot be reliably recovered from the source judgment.

4.3 Strict Provenance Patch

To reduce unsupported precedent identifiers flagged in the human review, we release an optional source-grounded patch. The audit covered 45,016 posthoc-complete judgments and 85,829 precedent Rules. Deterministic matching selected 41,731 Rules whose case name or citation could not be located reliably. GPT-OSS-120B then reviewed candidates using the judgment text, allowing abbreviations and OCR variants. Valid decisions were obtained for 24,824 candidates, while 16,907 remained unresolved and unchanged. The actionable patch changes 26,012 identifier fields across 12,476 Rules/Authorities entries. It provides a stricter extractive view, not external verification of legal correctness: absence from the focal judgment is not proof that an identifier is legally false.

4.4 LLM Screening and Annotator Agreement

We also run a full-corpus rubric screen with GPT-OSS-120B using a prompt analogous to the human rubric. At the aggregate component level, both assessments identify Rules/Authorities coverage as the lowest-scoring dimension: the LLM-screen mean is 4.207, with 24.3% below 4. Across all dimensions, the screen’s overall mean is 4.733, 78.8% of ratings are 5, and 4.9% are below 4. These aggregate tendencies do not imply case-level agreement.

Table 5 separates two questions. First, human exact agreement is 79.7%, within-one agreement is 96.3%, and mean absolute error (MAE) is 0.243. Ordinal Krippendorff’s α is moderate at 0.401, with a wide 95% interval of [0.120, 0.629], and should be interpreted alongside the ceiling effect in Table 3 and the limited 30-case overlap. A2 scored more generously and tightly than the other annotators (Appendix B). However, each annotator pair rated a different set of 10 shared cases, so annotator and item effects cannot be fully separated.

Second, GPT-OSS-120B is useful for triage but not as a replacement for human review. Within-one agreement is 94.4% against human means and 94.3% against human minima, but rank correlation is low ($\rho = 0.177$) and below-4 detection is weak ($F_1 = 0.131$ against human means). We therefore use it only as a corpus-health signal.

5 Supported Downstream Tasks and Baselines

The following baselines illustrate how the structured fields can be reused. Unless otherwise noted, LLM baselines use task-specific zero-shot prompts with temperature 0 and structured outputs, and retrieval or supervised baselines are fit separately within the IN-Abs and Supreme Court subsets. Tasks 2–4 use FIRAC-derived machine-generated targets, so their scores measure agreement with one reference view and do not establish unique legal correctness. In particular, Qwen3.6 may benefit when evaluated against Issues, Rules/Authorities, and outcome labels produced by the same model. Accordingly, cross-system comparisons report benchmark performance against the released targets and do not assess legal correctness. The results are reusable starting points, not task-specific state-of-the-art claims.

5.1 Task 1: Headnote-Style Structured Summarisation

We test whether FIRAC annotations can be rendered as legal headnotes. The input to annotation is the full judgment. Qwen3.6 produces the FIRAC structure, after which a deterministic renderer, with no additional LLM generation, writes a headnote-style summary. The renderer follows the IN-Abs headnote shape: Facts first, then an issue statement, followed by a “Held” section containing Conclusion, Rules/Authorities, and Analysis. Following Shukla et al. (2022), we also report an oracle length-controlled diagnostic variant that uses the gold reference summary’s per-case word count and performs section-aware unit selection under that budget. Because it uses reference length, it is not directly deployable.

We evaluate on the 100-case IN-Abs test split, the only released split with reference headnotes. R-1, R-2, and R-L are ROUGE-1, ROUGE-2, and ROUGE-L F1, measuring unigram, bigram, and longest-common-subsequence overlap with the reference. Scores are averaged over the 100 cases.

Table 4: LLM-judge rubric-score distribution over the full corpus.

Component	Dimension	Ratings (%)						Mean [95% CI]
		1	2	3	4	5	< 4	
Facts	Correctness	0.0	0.9	4.0	3.4	91.7	4.9	4.858 [4.854, 4.863]
	Coverage	0.0	0.1	3.9	39.1	57.0	4.0	4.529 [4.523, 4.534]
Issues	Correctness	0.0	0.1	0.8	3.1	96.1	0.9	4.951 [4.948, 4.953]
	Coverage	0.0	0.2	4.8	7.9	87.1	5.0	4.819 [4.815, 4.824]
Rules/Authorities	Correctness	0.2	0.8	2.4	9.8	86.8	3.4	4.823 [4.819, 4.828]
	Coverage	0.1	1.4	22.8	29.0	46.7	24.3	4.207 [4.199, 4.215]
Analysis	Correctness	0.1	0.4	0.5	10.8	88.3	0.9	4.869 [4.865, 4.872]
	Coverage	0.0	0.2	1.7	30.1	68.0	1.9	4.657 [4.652, 4.662]
Conclusion	Correctness	0.3	0.5	1.0	2.3	95.9	1.9	4.929 [4.925, 4.933]
	Coverage	0.2	0.3	1.1	27.6	70.8	1.6	4.685 [4.680, 4.690]
Overall		0.1	0.5	4.3	16.3	78.8	4.9	4.733 [4.730, 4.735]

Table 5: Human inter-annotator agreement and human-LLM alignment. Human-pair rows use case-criterion ratings from 30 double-annotated cases, whereas human-LLM rows compare the LLM with available human ratings over 120 cases.

Comparison	Exact (%)	Within 1 (%)	MAE	ρ	Ordinal α [95% CI]	$F_1(< 4)$	AUPRC
Human-human overall	79.7	96.3	0.243	0.405	0.401 [0.120, 0.629]	0.533	—
A1-A2	80.0	94.0	0.270	0.249	0.178 [-0.118, 0.555]	0.364	—
A1-A3	80.0	96.0	0.240	0.647	0.643 [0.183, 0.848]	0.667	—
A2-A3	79.0	99.0	0.220	0.384	0.240 [-0.027, 0.488]	0.500	—
LLM-human mean	68.9	94.4	0.358	0.177	—	0.131	0.102
LLM-human minimum	69.7	94.3	0.375	0.177	—	0.123	0.103

We also report BERTScore F1.⁴ Prior systems are copied from Shukla et al. (2022) for reference and were not rerun.

Table 6 shows that FIRAC annotations can be deterministically rendered into headnotes with competitive reference-overlap scores. The oracle length-controlled variant obtains the highest reported ROUGE-1, while BERTScore is comparable to the strongest reported systems. Lower ROUGE-L than the abstractive baselines may partly reflect the reorganisation of content. The expert component assessment in Table 3 evaluates source-grounded correctness and coverage of the underlying structure. We did not conduct a direct expert evaluation of rendered-summary fluency or pairwise preference. Supreme Court summarisation scores are not reported because verified reference headnotes are not available for that subset.

5.2 Task 2: Issue Identification

Issue identification tests whether Facts provide useful inputs for issue spotting. Given only Facts, the system must generate one to five dispositive legal questions without answering them or predict-

Table 6: Headnote-style summarisation on the IN-Abs test set.

System	R-1	R-2	R-L	BERT-F1
FIRAC headnote	0.497	0.240	0.243	0.854
FIRAC headnote (capped)	0.548	0.245	0.262	0.854
DSDR	0.485	0.222	0.270	0.848
CaseSummarizer	0.454	0.229	0.279	0.843
SummaRunner	0.493	0.255	0.274	0.849
Gist	0.471	0.238	0.308	0.842
BART-MCS	0.495	0.249	0.330	0.851
BERT-BART-MCS	0.487	0.243	0.329	0.853
Legal-Pegasus	0.488	0.252	0.341	0.851
Legal-LED	0.471	0.235	0.332	0.856

Prior rows are from Shukla et al. (2022), Table 4, and were not rerun.

ing the outcome. FIRAC Issues are the references. We evaluate on the 100-case IN-Abs test set and the 1,441-case temporal Supreme Court test set. Retrieval indexes contain 7,023 and 36,449 task-eligible training cases, respectively. In Table 7, #I is the mean number of predicted issues, Sem. P/R/F1 are soft semantic-matching precision, recall, and F1, and Tok. F1 is token F1.

We compare two retrieval controls with the generative systems. The Jaccard nearest-neighbour (NN) baseline copies all issues from the training case whose Facts have the highest token-set Jaccard similarity. The dense NN baseline embeds Facts with all-mpnet-base-v2, retrieves the near-

⁴BERTScore uses bert-score 0.3.4 with English roberta-large.

Table 7: Issue identification from Facts. Qwen3.6-27B also generated the reference annotations.

System	#I	Sem. P	Sem. R	Sem. F1	Tok. F1
<i>IN-Abs</i> ($n = 100$)					
Jaccard NN	1.59	.580	.580	.579	.354
MPNet NN	1.84	.590	.610	.598	.338
Qwen3.6-27B	2.19	.736	.782	.757	.550
Qwen3-32B	2.44	.696	.753	.721	.519
Mistral 24B	2.70	.706	.781	.739	.539
SaulLM-54B	1.11	.760	.727	.742	.496
<i>Supreme Court</i> ($n = 1,441$)					
Jaccard NN	1.40	.609	.602	.604	.339
MPNet NN	1.51	.612	.609	.609	.336
Qwen3.6-27B	2.41	.716	.764	.738	.514
Qwen3-32B	2.73	.679	.734	.704	.487
Mistral 24B	2.88	.690	.757	.721	.511
SaulLM-54B	1.08	.730	.698	.712	.441

est training case by cosine similarity, and copies its issues. All saved predictions are rescored consistently using paraphrase-MiniLM-L6-v2, independent of the MPNet retrieval encoder. Semantic precision averages each prediction’s best cosine match to a reference. Semantic recall averages each reference’s best match to a prediction. Semantic F1 is their harmonic mean. Scores are macro-averaged across cases. Token F1 takes, for each reference issue, the highest bag-of-token F1 against any predicted issue, then averages across references and cases.

Table 7 shows that generative systems substantially outperform both retrieval controls. Qwen3.6-27B obtains the strongest semantic and token F1 on both corpora, reaching 0.757 semantic F1 on IN-Abs and 0.738 on Supreme Court cases. Independent models remain well above retrieval: Mistral and SaulLM are not from the annotation model family, yet both achieve strong semantic F1. The systems also show a precision–coverage trade-off. SaulLM has the highest semantic precision but generates fewer issues and therefore has lower recall. Qwen3.6 and Mistral generate more issues and achieve stronger coverage. MPNet retrieval modestly improves semantic F1 over Jaccard, especially on IN-Abs, but not token F1, suggesting better conceptual rather than lexical matching. Because Qwen3.6-27B generated the reference Issues, its lead should not be read as independent evidence of annotation quality or as ruling out alternative formulations.

5.3 Task 3: Authority Recommendation

Authority recommendation evaluates whether a system can predict legal materials likely to govern an appeal based on a fact pattern and issue statement. The system ranks up to ten statutes, provi-

Table 8: Authority recommendation from Facts and Issues. Default rows use reference FIRAC Issues, whereas *pred. I* rows use Mistral-predicted Issues.

System	#A	P@1	R@5	nDCG@5	Tok. F1
<i>IN-Abs</i> ($n = 100$)					
Lexical BM25, top 5	5.00	.040	.028	.035	.324
Dense retrieval, top 10	10.00	.190	.064	.104	.417
Qwen3.6-27B	9.62	.330	.140	.214	.560
Qwen3-32B	9.65	.340	.120	.192	.510
Mistral 24B	9.90	.510	.185	.293	.521
SaulLM-54B	6.25	.230	.078	.126	.442
Mistral, pred. I	9.96	.420	.160	.250	.500
Mistral, pred. I + select	9.98	.310	.122	.188	.478
<i>Supreme Court</i> ($n = 1,441$)					
Lexical BM25, top 5	5.00	.031	.013	.020	.288
Dense retrieval, top 10	10.00	.201	.065	.108	.422
Qwen3.6-27B	9.84	.437	.163	.266	.521
Qwen3-32B	9.89	.364	.128	.212	.495
Mistral 24B	9.94	.366	.144	.232	.475
SaulLM-54B	6.29	.206	.064	.109	.358
Mistral, pred. I	9.96	.307	.116	.189	.456
Mistral, pred. I + select	10.00	.332	.113	.186	.480

sions, precedents, principles, or other authorities and the target is the FIRAC Rules/Authorities list. The main rows in Table 8 use reference FIRAC Issues. The end-to-end ablation substitutes Mistral-predicted Issues. We evaluate on the same 100 IN-Abs cases and 1,441 Supreme Court cases as above, with 7,023 and 36,444 eligible training cases. In the table, #A is the mean number of predicted authorities, P@1 is precision at 1, R@5 is recall at 5, nDCG@5 is normalized discounted cumulative gain at 5, and Tok. F1 is token F1.

We compare lexical, dense, retrieve-then-select, and unconstrained generative systems. The lexical baseline ranks training-set authorities using a lightweight BM25-style relevance of their associated case material. The dense baseline retrieves 20 training cases using MPNet Facts embeddings, aggregates their authorities by case similarity and rank, and returns the top ten. In the retrieve-then-select setting, Mistral receives its predicted Issues and the top 100 dense-retrieval candidates, and must select at most ten while copying identifiers exactly. Invalid outputs after retries are scored as empty.

For this illustrative benchmark, we use deliberately conservative string matching: each authority is represented by concatenating its reference, provision, case name, and citation fields, followed by lowercasing and alphanumeric normalization. P@1, R@5, and nDCG@5 count only identical normalized representations, so we interpret them as strict lower bounds on agreement with the dataset’s authority identifiers. Average-best token F1 exposes partial overlap, such as a correct statute with a missing provision or a case-name variant.

Table 9: Facts-only binary legal judgment prediction. Heterogeneous other outcomes are excluded.

System	Accuracy	Macro-F1	MCC
<i>IN-Abs</i> ($n = 98$; 48 allowed / 50 dismissed)			
Majority	.510	.338	.000
Qwen3.6-27B	.663	.657	.332
Mistral 24B	.561	.557	.121
Saul-7B	.490	.375	-.005
SauLLM-54B	.480	.449	-.056
TF-IDF + LR	.663	.663	.326
InLegalBERT	.622 $\pm$.037	.621 $\pm$.037	.246 $\pm$.076
<i>Supreme Court</i> ($n = 1,356$; 1,015 allowed / 341 dismissed)			
Majority	.749	.428	.000
Qwen3.6-27B	.694	.663	.387
Mistral 24B	.665	.631	.318
Saul-7B	.741	.450	.031
SauLLM-54B	.454	.453	.133
TF-IDF + LR	.733	.629	.260
InLegalBERT	.770$\pm$.012	.700$\pm$.011	.400$\pm$.020

Table 8 shows that dense retrieval substantially improves over lexical retrieval, but unconstrained LLM recommendation remains stronger. Mistral leads exact ranking metrics on IN-Abs, while Qwen3.6-27B leads on Supreme Court and obtains the strongest token F1 on both corpora. SaulLM supplies a legal-domain baseline, but does not outperform the general models. Replacing reference Issues with predicted Issues lowers Mistral’s P@1 from .510 to .420 on IN-Abs and from .366 to .307 on Supreme Court, demonstrating measurable upstream error propagation. Retrieve-then-select is mixed and does not consistently improve over unconstrained Mistral. The top-100 pool recalls only .158 of IN-Abs and .219 of Supreme Court reference authorities, making candidate recall the principal bottleneck.

This strict procedure likely produces false negatives: the same authority may appear through alternative case names, reporter citations, statute variants, abbreviations, reordered terms, or partially populated fields. For Qwen3.6-27B, token-F1 $\geq .8$ near-matches account for another 22.3% of IN-Abs and 14.9% of Supreme Court authorities beyond exact matches. More realistic entity-level evaluation would require canonical legal record normalisation and entity-linking, which is out of the scope of this publication, but remains future work. The modest strict-match scores therefore reflect both recommendation difficulty and identifier variation, and should be read alongside the human Rules/Authorities assessment and the optional strict provenance patch in Section 4.3.

5.4 Task 4: Binary Legal Judgment Prediction

Finally, FIRAC-APPEALS supports leakage-aware legal judgment prediction (LJP). We frame the

benchmark as conditional binary LJP: after eligibility for the two-label task is established, the system predicts allowed versus dismissed using only Facts. Heterogeneous other cases are excluded. Source-specific splits contain 6,877/98 IN-Abs train/test cases and 34,697/1,356 temporal Supreme Court train/test cases. Macro-F1 is primary because the Supreme Court test set is imbalanced. Malformed outputs count as errors. The target is the normalized simplified label emitted with Qwen3.6’s FIRAC Conclusion, so this task also inherits annotation-model dependence and the coarse mapping of partly allowed cases to allowed.

For LJP, LLMs return a label plus a short prospective rationale. We also train supervised baselines: InLegalBERT (Paul et al., 2023), reported as mean $\pm$ sample standard deviation over three seeds, and class-balanced lexical logistic regression (TF-IDF+LR). Appendix C gives training details and coefficient diagnostics.

Table 9 shows that Facts-only binary LJP remains difficult but informative. On IN-Abs, TF-IDF+LR is strongest by macro-F1 (.663), approximately matching Qwen3.6-27B (.657), while InLegalBERT reaches .621 $\pm$.037 on a small 98-case test set. On temporal Supreme Court cases, InLegalBERT is strongest, reaching .700 $\pm$.011 macro-F1 and .400 $\pm$.020 MCC. The majority baseline has high accuracy (.749) but macro-F1 .428 and MCC 0, demonstrating why accuracy alone is unsuitable. TF-IDF+LR improves macro-F1 to .629 despite lower accuracy, showing meaningful minority-class prediction.

Qwen3.6-27B remains the strongest zero-shot LLM, while independent supervised models match or exceed it. This shows that useful predictors need not share the annotation model, but it does not remove dependence on Qwen-derived labels. Legal-domain instruction tuning also does not automatically transfer to prospective LJP: Saul-7B largely reproduces the Supreme Court majority distribution, and SauLLM-54B performs poorly. The lexical coefficients in Appendix C show that procedural and narrative framing remains predictive. We therefore cannot apportion performance between substantive fact-pattern reasoning and such shortcuts, and do not present outcome accuracy as evidence of legal reasoning.

6 Future Work

Future extensions include external canonical authority resolution beyond the source-grounded patch in Section 4.3, an independent-generator pilot, paraphrase-level train–test deduplication, and direct expert preference evaluation of rendered headnotes. We also plan to extend the pipeline to High Court judgments and to develop judgment-generation quality metrics beyond final outcomes or surface overlap. FIRAC segmentation is suited to this because issue framing, authority grounding, rule application, and conclusion fidelity can be evaluated separately.

7 Conclusion

FIRAC-APPEALS turns Indian appellate judgments into reusable, machine-generated FIRAC components: Facts, issue statements, rules and authorities, analysis, conclusions, metadata, and outcome labels. The resource shows how lightweight legal structure can make public judgment collections more useful for summarisation, authority grounding, issue spotting, outcome prediction, and future reasoning-trace evaluation. Our experiments provide evidence of usability and baseline measurements rather than task-specific state-of-the-art or legal-correctness claims. By exposing intermediate components and their known weaknesses, FIRAC-APPEALS provides a foundation for more inspectable legal NLP research.

Limitations

The resource covers one court, one jurisdiction, and English-language judgments, so findings should not be generalized to Indian High Courts, other jurisdictions, or multilingual settings. All FIRAC references come from one annotation model and may reproduce its omissions, framing choices, and systematic errors. Expert assessment covers 120 stratified cases, with 30 double annotations and moderate, uncertain chance-corrected agreement. It evaluates source faithfulness and coverage, not authoritative legal correctness. The full-corpus LLM screen has weak case-level alignment with humans and is only a triage signal.

Facts are reconstructed retrospectively from completed judgments rather than prospectively authored case files. Exact normalized duplicates and direct outcome phrases were checked, but paraphrase-level overlap was not ruled out. Procedural posture and lower-court dispositions may

still provide indirect outcome cues. The temporal Supreme Court test also differs substantially in outcome distribution from training. The three-way outcome taxonomy compresses partial relief and mixed dispositions. Binary LJP further excludes other. The full Conclusion should therefore be used whenever disposition nuance matters. Corpus-level compression statistics do not establish preservation of nuance in particular legal domains or high-stakes cases, and no direct human preference study of the rendered headnotes was conducted. Rules/Authorities remains the weakest component, and the source-grounded patch does not externally verify legal correctness.

Ethical Considerations

Public availability of judgments does not make every downstream use harmless. The texts can contain names and sensitive details concerning offences, family matters, employment, health, finances, and other disputes, and they reflect institutional, procedural, socioeconomic, and historical biases. Machine-generated summaries or predictions can amplify those biases, omit qualifications, invent or misstate authorities, and create automation bias if presented as professional legal analysis. FIRAC-APPEALS is intended for research on structured summarisation, retrieval, audit, and carefully scoped prediction. It is not intended for autonomous adjudication, legal advice, litigant or judge profiling, risk scoring, or replacement of source checking and professional review. Mitigations include source provenance, released prompts and artifacts, explicit quality limitations, a source-grounded authority view, and retention of the original judgment for verification. Users remain responsible for privacy, licensing, jurisdiction-specific obligations, and the effects of any deployment.

Acknowledgments

We thank the legal annotators for their careful evaluation work, insightful comments, and detailed feedback on the FIRAC annotations. We also thank the maintainers of the public Indian legal NLP and Supreme Court judgment resources on which this work builds.

References

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. [Can GPT-3 perform statutory](#)

- reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23*, pages 22–31, New York, NY, USA. Association for Computing Machinery.
- Joy Bose. 2026. [Falkor-IRAC: Graph-constrained generation for verified legal reasoning in Indian judicial AI](#). *arXiv preprint arXiv:2605.14665*. Submitted on 14 May 2026; last revised 15 May 2026 (v2).
- Daniel Braun. 2024. [I beg to differ: How disagreement is handled in the annotation of legal machine learning data sets](#). *Artificial Intelligence and Law*, 32(3):839–862.
- Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. [MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302, Singapore. Association for Computational Linguistics.
- Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. [LawBench: Benchmarking legal knowledge of large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. [LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.
- Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Zhuang Li, and Adnan Trakic. 2025. [Automating IRAC analysis in Malaysian contract law using a semi-structured knowledge base](#). *Artificial Intelligence and Law*.
- Manuj Malik, Zheng Zhao, Marcio Fonseca, Shrisha Rao, and Shay B. Cohen. 2024. [CivilSum: A dataset for abstractive summarization of Indian court decisions](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2241–2250, New York, NY, USA. Association for Computing Machinery.
- Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. [ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.
- Masha Medvedeva and Pauline McBride. 2023. [Legal judgment prediction: If you are going to do it, do it right](#). In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84, Singapore. Association for Computational Linguistics.
- Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. [Rethinking the field of automatic prediction of court decisions](#). *Artificial Intelligence and Law*, 31:195–212.
- Shubham Kumar Nigam, Tanmay Dubey, Govind Sharma, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. [LegalSeg: Unlocking the structure of Indian legal judgments through rhetorical role classification](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1129–1144, Albuquerque, New Mexico. Association for Computational Linguistics.
- Vedant Parikh, Vidit Mathur, Parth Mehta, Namita Mittal, and Prasenjit Majumder. 2021. [LawSum: A weakly supervised approach for Indian legal document summarization](#). *arXiv preprint arXiv:2110.01188*.
- Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. [Pre-trained language models for the legal domain: A case study on indian law](#). In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23*, pages 187–196, New York, NY, USA. Association for Computing Machinery.
- Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. [Legal case document summarization: Extractive and abstractive methods and their evaluation](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only. Association for Computational Linguistics.
- Michael B. W. Sinclair. 2003. [What is the “R” in “IRAC”?](#) *New York Law School Law Review*, 46(3):457–474.
- Tracy L. Turner. 2015. [Flexible IRAC: A best practices guide](#). *Legal Writing: The Journal of the Legal Writing Institute*, 20:233–273.
- Bin Wei, Yaoyao Yu, Leilei Gan, and Fei Wu. 2025. [An LLMs-based neuro-symbolic legal judgment prediction framework for civil cases](#). *Artificial Intelligence and Law*.
- Yue Zhang, Zhiliang Tian, Shicheng Zhou, Haiyang Wang, Wenqing Hou, Yuying Liu, Xuechen Zhao, Minlie Huang, Ye Wang, and Bin Zhou. 2025. [RLJP](#):

A Released Dataset Schema

Table 10 summarises the public Parquet row schema. The schema is intentionally compact: it keeps the fields needed for downstream tasks and inspection, while leaving run logs, raw prompts, rubric judgments, and evaluation artifacts in the reproducibility package.

B Human Evaluation Protocol and Diagnostics

The released reproducibility package contains the exact Label Studio interface, component-specific rubric, sanitized annotations, assignment manifest, and reporting code. Before evaluation, annotators were given 10 non-evaluation practice cases drawn from training data. Main ratings were completed independently and were not adjudicated. Each annotator pair shared 10 different cases. Consequently, pairwise severity comparisons are partly confounded with item choice. For Facts, the rubric treats outcome leakage as one possible correctness defect rather than a separate rating dimension.

C LJP Training and Diagnostics

For InLegalBERT, we fine-tune law-ai/InLegalBERT independently on each source’s Facts. For each seed, we use a stratified 90/10 train/validation split, truncate inputs to 512 tokens, dynamically pad each batch, and train for three epochs with batch size 32, AdamW, learning rate 2×10^{-5} , weight decay 0.01, 10% linear warm-up, gradient clipping at 1.0, bfloat16 mixed precision, and class-weighted cross-entropy. We select the checkpoint with the highest validation macro-F1. Seeds 20260604, 20260605, and 20260606 affect the train/validation split, classifier-head initialization, minibatch order, dropout, and GPU training. An eight-epoch run with patience-two early stopping did not improve aggregate test macro-F1, so we retain the simpler three-epoch protocol.

For TF-IDF+LR, we concatenate word TF-IDF features (1–2 grams, maximum 150k) and character-within-word features (3–5 grams, maximum 150k), use sublinear term frequency, and train a class-balanced logistic regression with $C = 1$,

liblinear, and a fixed seed. The model is trained separately on the same source-specific binary training sets as InLegalBERT.

The TF-IDF coefficients make the baseline useful as both a competitive non-neural model and a diagnostic for residual lexical cues. Character features reinforce similar stems, especially variants of *argu-*, *petition-*, and *contend-*, but are less interpretable and are not printed. The word features suggest that procedural and narrative framing carries outcome signal: petitioner or appellant argument language is associated with dismissal, while language describing court holdings, reliance, and respondent arguments is associated with allowance. Some terms may encode case type, procedural posture, or source-specific writing conventions rather than substantive legal reasoning, and recurrent phrasing from the common FIRAC annotation pipeline may make such cues especially learnable. We found no overlapping case identifiers or exact normalized Facts between train and test. Direct phrases such as “appeal allowed” or “appeal dismissed” were absent from the IN-Abs test set and very rare in the Supreme Court test set. This reduces the likelihood of trivial label leakage, but does not eliminate indirect outcome cues, particularly descriptions of lower-court dispositions or procedural posture. We therefore quantify and disclose this possible shortcut rather than presenting supervised classifier performance as evidence of legal reasoning.

Table 10: Public row schema. List fields are ordered as generated. Empty strings or empty lists indicate unavailable source information.

Field	Type	Purpose
case_id	string	Stable identifier for the released row.
source_dataset	string	Source subset, currently <code>in_abs</code> or <code>supreme_court</code> .
source_case_id	string	Identifier inherited from the upstream source where available.
split	string	Released split membership, <code>train</code> or <code>test</code> .
judgment_text	string	Full judgment text used as input to annotation and downstream task construction.
reference_summary	string	Upstream IN-Abs headnote/reference summary, empty for SC rows.
simplified_outcome_label	string	Coarse ML label, <code>allowed</code> , <code>dismissed</code> , or <code>other</code> . Not a substitute for the full disposition.
metadata	struct	Cleaned case metadata: title, court, decision date, citations, case numbers, parties, judges/bench, judgment author, impugned court, lawyers, disposal nature, available languages, and source year where available.
firac.facts	list[string]	Factual and procedural background, excluding holdings, reasoning, and outcome language where possible.
firac.issues	list[string]	Dispositive legal questions framed without answers.
firac.rules	list[struct]	Authorities and legal principles considered relevant to the appeal. Fields are detailed in Table 11.
firac.analysis	list[struct]	Issue-level application, with issue, ordered reasoning steps, and local conclusion.
firac.conclusion	struct	Disposition object with original-style label, fuller explanatory text, and simplified_label.

Table 11: Fields in each `firac.rules` entry. The fields support both exact identifier matching and more forgiving text-based inspection.

Field	Type	Purpose
type	string	Authority category, e.g. precedent, statute, constitutional provision, regulation, legal principle, doctrine, or other.
reference	string	Human-readable authority reference, such as a statute title, case reference, doctrine, or principle.
provision	string	Section, article, rule, clause, or other provision-level locator where applicable.
case_name	string	Case name for precedent entries, separated from citation to support entity-level normalization.
citation	string	Reporter, neutral, or other citation string where available or recoverable.
treatment_type	string	How the judgment appears to use the authority, such as <code>relied on</code> , <code>referred</code> , <code>distinguished</code> , <code>overruled</code> , or <code>not applicable</code> .
relevance	string	Short explanation of why the authority matters for the appeal.
used_by_court	boolean	Whether the authority is represented as used by the court rather than merely raised by a party or inferred as background.

Table 12: Annotator score distributions. Mean, SD, and variance are computed across the 50 case-level means. Score 5 and < 4 are percentages of the 500 individual rubric scores.

Annotator	Mean	SD	Variance	Score 5 (%)	< 4 (%)
A1	4.622	0.517	0.267	76.2	10.0
A2	4.962	0.070	0.005	96.8	0.6
A3	4.744	0.290	0.084	78.2	2.6

Table 13: Highest-weight word features in the TF-IDF+LR binary LJP baseline. Negative coefficients indicate allowed, whereas positive coefficients indicate dismissed. These are global model coefficients, not causal explanations or per-case attributions.

Subset	Allowed features	Dismissed features
IN-Abs	relying on (-1.09), holding that (-1.09), relying (-.98), detention (-.85), holding (-.78)	petitioner (1.83), the petitioner (1.79), appellant argued (1.38), arguing that (1.18), leave petition (1.16)
Supreme Court	respondents argued (-2.25), respondent argued (-2.17), court relied (-2.15), reasoning (-1.89), holding that (-1.85)	the petitioner (3.97), appellant argued (3.66), appellant did (2.95), petitioner (2.92), filed special (2.72)