

Structured Legal Document Generation in India: A Model-Agnostic Wrapper Approach with VidhikDastaavej

Shubham Kumar Nigam^{1,2*†} Balaramamahanthi Deepak Patnaik^{1*}

Noel Shallum³ Kripabandhu Ghosh⁴ Arnab Bhattacharya¹

¹Indian Institute of Technology Kanpur, India; ²University of Birmingham Dubai, UAE

³Symbiosis Law School Pune, India; ⁴IISER Kolkata, India

{shubhamkumarnigam, bdeepakpatnaik2002, noelshallum}@gmail.com

kripaghosh@iiserkol.ac.in, arnabb@cse.iitk.ac.in

Abstract

Automating legal document drafting can improve efficiency and reduce the burden of manual legal work. Yet, the structured generation of private legal documents remains underexplored, particularly in the Indian context, due to the scarcity of public datasets and the complexity of adapting models for long-form legal drafting. To address this gap, we introduce *VidhikDastaavej*, a large-scale, anonymized dataset of private legal documents curated in collaboration with an Indian law firm. Covering 133 diverse categories, this dataset is the first resource of its kind and provides a foundation for research in structured legal text generation and Legal AI more broadly. We further propose a Model-Agnostic Wrapper (MAW), a two-stage generation framework that first plans the section structure of a legal draft and then generates each section with retrieval-based prompts. MAW is independent of any specific LLM, making it adaptable across both open- and closed-source models. Comprehensive evaluation, including lexical, semantic, LLM-based, and expert-driven assessments with inter-annotator agreement, shows that the wrapper substantially improves factual accuracy, coherence, and completeness compared to fine-tuned baselines. This work establishes both a new benchmark dataset and a generalizable generation framework, paving the way for future research in AI-assisted legal drafting.

Keywords: Legal Document Generation, Legal Drafting Automation, Structured Text Generation, Long-Form Text Generation, Model-Agnostic Wrapper (MAW), Legal Language Resources, Dataset Annotation, Expert Evaluation, Human-in-the-loop, Legal Document Planning, Legal Text Benchmarking, Controlled Text Generation

1. Introduction

Automating legal document generation can significantly improve efficiency and accessibility in legal workflows. Although LLMs have been widely used for legal tasks such as prediction of judgments, summarization of cases, and retrieval, their application to the generation of private legal documents remains underexplored, particularly in the Indian legal domain. The primary challenge lies in the confidentiality of private legal documents, which limits publicly available training data.

To address this, we introduce *VidhikDastaavej*, a novel anonymized dataset of private legal documents, collected in collaboration with Indian legal firms. The name *VidhikDastaavej* is derived from the Hindi words “Vidhik” (legal) and “Dastaavej” (documents), reflecting its focus on legal document automation. This dataset serves as a valuable resource for training and evaluating structured legal text generation models, while ensuring compliance with ethical and privacy standards.

To further complicate matters, the landscape of large language models is evolving at a rapid pace, with new models being released frequently. In such a scenario, methods that rely on task-specific

supervised fine-tuning (SFT) quickly become outdated or impractical, especially when a newer, more powerful model is introduced shortly after. Moreover, most end-users, such as legal practitioners or developers working with proprietary or custom-deployed models, may not have the resources to retrain or fine-tune large models. In some cases, users may prefer to keep their model private or operate within hardware constraints that prevent full-scale training. This raises an urgent need for model-agnostic approaches that can adapt seamlessly across different LLMs without requiring architectural modifications or extensive retraining. To overcome this challenge, we propose a lightweight and scalable *Model-Agnostic Wrapper (MAW)* for structured legal document generation. The wrapper decouples the generation process from any particular model by adopting a two-stage workflow: first generating section titles from document instructions, followed by iterative content generation for each section. This structure-then-generate strategy promotes coherence, reduces hallucinations, and ensures factual alignment, all while remaining compatible with any base LLM, whether open-source, commercial, or privately hosted. This flexibility makes our approach particularly valuable for real-world legal applications where model diversity and resource constraints are the norm.

*These authors contributed equally to this work

†Corresponding author

For rigorous evaluation, we introduce expert-based assessment, where legal professionals review generated documents based on factual accuracy (adherence to legal instructions) and completeness and comprehensiveness (coverage of all essential details) between 1–10 (Irrelevant–Relevant) Likert scale. This ensures a robust evaluation beyond standard lexical and semantic metrics, addressing the complexity of legal drafting.

Additionally, we provide an interactive Human-in-the-Loop (HITL) Document Generation System, enabling users to input document types, customize sections, and generate structured legal drafts. To enhance reproducibility, we have made the `VidhikDastaavej` dataset, models, and codes available through a GitHub repository[†].

To the best of our knowledge, this is the first work in the Indian legal domain focusing on automated private legal document generation. Our key contributions include:

1. `VidhikDastaavej` Dataset: A novel, anonymized dataset of private legal documents for structured legal text generation.
2. Model-Agnostic Wrapper: A structured framework ensuring coherence, consistency, and factual accuracy in generated legal drafts.
3. Expert-Based Evaluation Metrics: Introduction of structured legal evaluation focusing on factual accuracy and completeness.
4. Human-in-the-Loop System: A user-friendly interface for structured legal document generation, supporting practical legal workflows.

This research lays the foundation for AI-assisted legal drafting in India, modernizing legal workflows while ensuring accuracy, consistency, and legal compliance.

2. Related Work

AI and NLP have advanced significantly in the legal domain, supporting applications such as judgment prediction (Medvedeva and McBride, 2023), legal case summarization (Ragazzi et al., 2024; Moro et al., 2024; Shukla et al., 2022), semantic segmentation (Moro and Ragazzi, 2022), and legal NER (Päis et al., 2021). These studies demonstrate the potential of AI systems to improve transparency, efficiency, and explainability in legal practice. In the Indian context, prior work has largely focused on public case judgments, particularly from the Supreme Court and High Courts, for retrieval, reasoning, and explainability tasks (Chalkidis et al., 2020). Datasets such as ILDC (Malik et al., 2021), PredEx (Nigam et al., 2024), and NyayaAnumana (Nigam et al., 2025)

support judgment prediction and rationale extraction, while rhetorical role labeling and segmentation tasks have been addressed using models ranging from CRF-BiLSTM (Bhattacharya et al., 2019) to hierarchical frameworks like HiCuLR (Santosh et al., 2024). These datasets facilitate training transformer-based models to enhance explainability and decision support systems for Indian legal texts (Nigam et al., 2022; Malik et al., 2022; Nigam et al., 2023), while recent studies have also addressed legal NER using large-scale pretrained models (Vats et al., 2023).

Globally, legal text generation has begun to gain traction, extending beyond judgment analysis to structured drafting tasks. Early studies explored controlled natural language drafting (Tateishi et al., 2019), segmentation-assisted contract generation (Tong et al., 2022), and text style transfer models (Li et al., 2021). Knowledge graph-based approaches have also been proposed to enforce coherence in drafting (Wei, 2024). Recent advances include retrieval-augmented and planning-driven methods for legal drafting and summarization: LexGenie (T.y.s.s et al., 2025a) generates structured multi-case reports for European Court of Human Rights (ECHR) law; CLERC (Hou et al., 2025) enables case retrieval and retrieval-augmented reasoning for U.S. law; LexKeyPlan (T.y.s.s and Hernandez, 2025) introduces anticipatory keyphrase planning to improve retrieval-grounded generation; and CoPER-Lex (T.y.s.s et al., 2025b) demonstrates event-centric content planning for case summarization. In the legislative domain, LexDrafter (Chouhan and Gertz, 2024) employs RAG for harmonized terminology drafting in EU legal texts. More recently, tools like LEGALSEVA (Pandey et al., 2024) and Legal DocGen (Patil et al., 2024) demonstrate efforts toward automated drafting of legal contracts and forms.

Despite these global developments, automation of *private* legal document drafting in India remains largely unexplored due to the unavailability of curated datasets. Our work fills this gap by introducing `VidhikDastaavej`, a large-scale anonymized dataset of diverse private legal documents, and proposing a structured, model-agnostic wrapper framework for coherent and factually accurate document generation. This positions our work at the intersection of dataset creation, structured generation methodology, and rigorous legal expert evaluation.

3. Problem Statement

The primary objective of this work is to develop a system that can automatically generate private legal documents based on specific user prompts or

[†]<https://github.com/ShubhamKumarNigam/VidhikDastaavej>

Metric	Train	Test
Number of documents	11,692	133
Number of unique categories	133	133
Avg # of words per document	5,798.61	7,464.62
Max # of words per document	98,607	81,233

Table 1: Dataset statistics for `VidhikDastaavej`.

situational inputs. Given an input x , which includes detailed instructions or contextual information, the task is to produce a legal document y that aligns with professional legal drafting standards in the Indian legal domain.

Formally, the problem can be defined as learning a function f such that: $y = f(x)$, where:

- x represents the user-provided prompt containing specific instructions, situational details, and any particular requirements for the legal document.
- y is the generated legal document that accurately reflects the content of x and is properly formatted and structured according to legal conventions.

The challenge lies in accurately mapping the input x to a coherent and contextually appropriate document y . This requires the system to understand and interpret complex legal language, terminologies, and document structures specific to the Indian legal context. The goal is to leverage LLMs to perform this mapping effectively, enabling the generation of high-quality legal documents that meet professional standards.

4. Dataset

To develop our automated legal document generation tool, we collaborated with an Indian legal firm to curate `VidhikDastaavej`, a novel, large-scale, anonymized dataset of private legal documents. This partnership granted access to a diverse collection of legal drafts that are not publicly available, ensuring that our dataset reflects real-world legal drafting practices in the Indian legal system. The dataset is designed not only as training material but also as a benchmark for the broader Legal AI community.

4.1. Dataset Composition and Diversity

The dataset encompasses a wide variety of documents, including License Agreements, Severance Agreements, Stock Option Agreements, Consulting Agreements, Asset Purchase Agreements, and more. By incorporating multiple document types, `VidhikDastaavej` captures diverse structures, terminologies, and drafting conventions in legal writing, moving beyond the traditional focus on case judgments seen in public legal datasets.

Table 1 provides an overview of the dataset statistics. `VidhikDastaavej` consists of 11,825 documents, with 11,692 used for training and 133 re-

served for testing. The dataset covers 133 distinct legal document categories in both training and testing, offering a broad representation of real-world legal drafts. A visual distribution of the top 15 most frequent document categories is presented in Figure 1. The complete category list is provided in the anonymous GitHub repository and dataset folder.

4.2. Annotation and Expert Curation

Since legal drafting requires domain expertise, we employed a multi-stage annotation and validation pipeline:

- **Document Categorization:** Initial categorization was performed using LLM-based classification (Mixtral) followed by validation from practicing legal experts. Each document was mapped to one of 133 categories. Experts verified the accuracy of categorization to avoid propagation of noisy labels.
- **Section-Level Structuring:** For training prompts, section headers were generated using LLaMA-3.1 models and then manually reviewed by experts to ensure alignment with professional drafting norms (e.g., inclusion of clauses, definitions, governing law sections).
- **Expert Instructions:** Experts were provided with clear guidelines on assessing legal soundness. They were asked to evaluate factual accuracy (whether the draft adhered to given instructions and contained no hallucinations) and completeness (coverage of mandatory sections and details). These were rated on a 1–10 Likert scale.
- **Inter-Annotator Reliability:** To ensure reliability, three independent legal experts reviewed the generated drafts, and inter-annotator agreement metrics (Fleiss’ κ , ICC, Krippendorff’s α) were computed. Results demonstrated strong agreement, validating the robustness of expert evaluations.

4.3. Data De-identification (NER-based Anonymization)

To comply with privacy regulations and ethical standards, all documents underwent rigorous anonymization. All documents were de-identified prior to any analysis or model training. We applied a Named Entity Recognition (NER) based redaction procedure using spaCy (`en_core_web_sm`). For each document, spaCy detects entity spans (e.g., PERSON, ORG, GPE, LOC, DATE) and we replace each detected span with a type placeholder of the form `[LABEL]`.

Formally, given a document x and the set of detected entity spans $\mathcal{E}(x) = \{(s_i, e_i, \ell_i)\}_{i=1}^n$, where (s_i, e_i) are character offsets and ℓ_i is the entity label, we produce a de-identified document \tilde{x} by replacing every substring $x[s_i : e_i]$ with `[ℓ_i]` while preserving

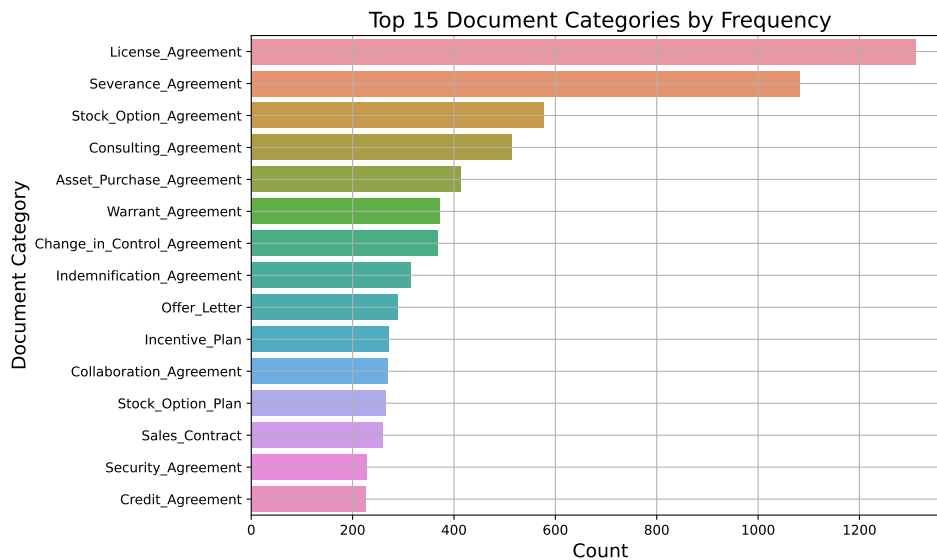


Figure 1: Top 15 Document Categories in the VidhikDastaavej Dataset based on Frequency.

all non-entity text unchanged. Table 2 shows an example output.

Human verification. Because automated NER-based redaction can miss identifiers depending on writing style and context, we conducted expert manual inspections on a subset of the corpus to check for residual personally identifying information. The inspected subset did not reveal remaining identifiers in the reviewed documents; however, we acknowledge that no automated method can guarantee perfect anonymization in all cases.

4.4. Significance of the Dataset

Unlike prior datasets that focus primarily on court judgments or narrow subsets of legal texts, VidhikDastaavej provides a broad, real-world representation of private legal documentation in India. This diversity allows models to learn structural conventions, drafting styles, and domain-specific vocabulary. The dataset serves as a foundational benchmark for the Legal AI community, supporting tasks such as document generation, section planning, and factual verification. By releasing both the anonymized corpus and the evaluation setup, we aim to foster reproducible research and enable further advancements in AI-assisted legal drafting.

5. Model-Agnostic Wrapper (MAW)

To improve long-form legal document generation, we introduce a Model-Agnostic Wrapper (MAW), a framework designed to integrate with any LLM for structured drafting. Legal documents require maintaining logical flow, coherence, and factual accuracy, which general-purpose LLMs often struggle

with when handling extended text generation.

The MAW employs a two-phase workflow (Figure 2) to ensure structured, contextually relevant content generation.

Phase 1: Section Title Generation. In the first phase, section titles are generated based on user input. The process begins with the user providing a document title and a brief description of the intended document. These inputs are passed to the chosen language model, which then generates a structured list of section titles. The generated section titles are displayed to the user, who can review and modify them, renaming, inserting new sections, or removing unnecessary ones before proceeding to content generation. Once the section titles are finalized, the process transitions to the next phase.

Phase 2: Section Content Generation. In the second phase, content is generated iteratively for each section. The workflow follows these steps:

1. For each section title, the model receives document title and description as additional context.
2. The model generates detailed section content along with a concise summary of the section.
3. The generated summary is stored in a vector database (ChromaDB) (Team, 2023) to facilitate contextual referencing.
4. During subsequent iterations, the vector database is queried for relevant section summaries, which are then incorporated into the LLM's context to enhance coherence and maintain logical document flow.
5. After generating content for all sections, the final document is refined and structured, ensuring clarity and coherence.

By adopting a two-phase workflow, we ensure that adequate time is dedicated to both section title generation and section content generation sep-

Power of Attorney
To All of whom, these presents shall come, I [PERSON] of [GPE] send Greetings
Whereas,
 1. Mr. [PERSON] shall appoint some fit and proper person to carry on acts for me and manage all my affairs.
 2. I nominate, constitute, and appoint my brother, Mr. [PERSON], as my true and lawfully appointed attorney (hereinafter called the Attorney) to act for me in the court of law for court proceedings in the matter of disputed joint property.
NOW THIS PRESENT WITNESSETH AS FOLLOWS:
 1. The attorney shall handle all the affairs with regard to court proceedings in the matter of disputed joint property.
 2. All the filings of applicants and suits in the court of law.
 3. All the appearances in the court proceedings.
 4. All the costs, expenses, and fees with regard to court proceedings.
 5. The fees to be paid to the lawyer appointed.
And I, Mr. [PERSON], undertake to ratify all the acts of the attorney or any agent appointed by him.
IN WITNESS WHEREOF, I set and subscribe my hand on [DATE].

[WORK_OF_ART] by within named.
 Mr. [PERSON] above named in the presence of:

1. _____ Mr. [PERSON]
 2. _____ Mr. [PERSON]

Table 2: This table illustrates a sample document after it has been anonymized.

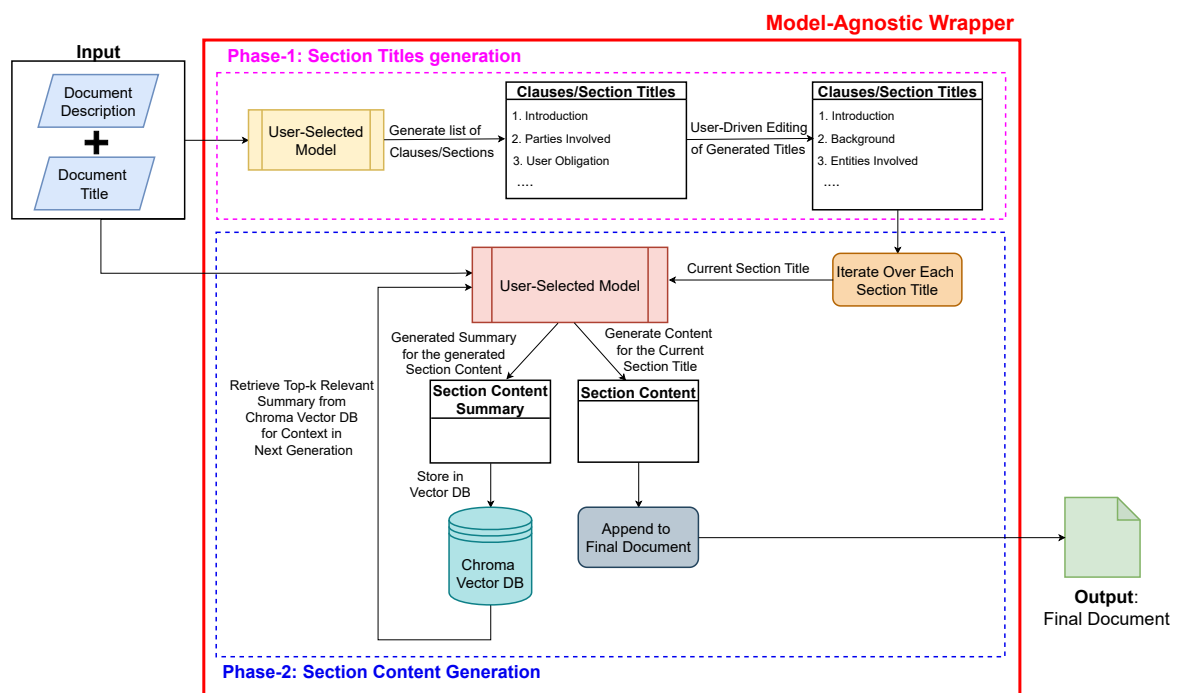


Figure 2: Wrapper flow diagram

arately, rather than attempting to generate both simultaneously. This separation allows for better coherence, logical structuring, and improved alignment between titles and their corresponding content, thereby enhancing the overall quality and readability of the generated document.

To demonstrate practical usability, we have developed a Human-in-the-Loop (HITL) Document Generation System that allows users to specify document types, refine sections, and interactively

generate drafts. Due to LREC submission policies, the complete user guide with detailed instructions and screenshots is not included at this stage, but will be provided in the camera-ready version as supplementary material.

6. Experimental Setup

To benchmark our pipeline's performance and assess our wrapper's effectiveness, we conducted

Category	Prompt
Development Agreement	Create a development, license, and hosting agreement between [ORG] and [ORG] LLC, effective as of [DATE], outlining the terms and conditions for the development, licensing, and hosting of [ORG], including [ORG] and [ORG], for the sale of [ORG] flights and other air transportation services through the [ORG] website. The agreement should include provisions for the definition of key terms, the scope of work, the schedule, the fees, the payment terms, the confidentiality obligations, the intellectual property rights, the warranties and disclaimers, the indemnification obligations, the limitation of liability, the insurance requirements, the dispute resolution procedures, the term and termination provisions, and the miscellaneous provisions. The agreement should also include exhibits for the specifications, the change request, the schedule, the fees, the relationship managers, the service level agreement, the non-disclosure agreement, and the escrow agreement
Purchase Agreement	Create a purchase agreement between [WORK OF ART] and Stacked Digital LLC, outlining the terms and conditions of the sale, including the purchase price, payment terms, delivery schedule, warranties, and any other relevant details necessary for a comprehensive agreement between the [CARDINAL] parties.

Table 3: Categories and Corresponding Prompts for Legal Document Generation

instruction tuning on various open-source models and compared them against GPT-4o.

6.1. Fine Tuning of Open-Source Models

We fine-tuned select open-source models while directly evaluating others without additional training. The instruction-tuned models include Qwen3-14B (Yang et al., 2025), LLaMA-3.1-8B-Instruct (Dubey et al., 2024), and Gemma-3-12B-It (Team et al., 2025) SFT to assess improvements in structured legal drafting.

For instruction tuning, we designed specialized prompts and instruction sets tailored to legal drafting. These instructions provided structured examples, ensuring that the models understood the nuances of different types of legal documents.

6.2. Benchmarking with GPT-4o

To assess the effectiveness of our instruction-tuned models and the Model-Agnostic Wrapper, we benchmarked performance against GPT-4o, a proprietary closed-source model. Unlike the open-source models, GPT-4o was not instruction-tuned but was used purely for inference. This comparison highlights the potential of fine-tuned open-source models as cost-effective alternatives for structured legal drafting, offering insights into whether instruction tuning can achieve performance comparable to commercial LLMs.

6.3. Hyperparameters

All experiments were conducted using the PyTorch framework integrated with Hugging Face Transformers. For SFT (Supervised Fine-Tuning), we used four NVIDIA H200 (Neysa) GPUs with 80GB of memory each. Mixed-precision training (fp16) was enabled to optimize memory and computational efficiency, and training progress was logged with Weights & Biases for effective monitoring.

We fine-tuned three instruction models, Qwen3-14B, Gemma-3-12B-It, and LLaMA-3.1-8B-Instruct, on the expanded dataset. Each model supported a maximum sequence length of 4500 tokens, allowing for long-context learning essential for legal drafting tasks.

The optimization was performed using the AdamW optimizer with a learning rate of 1×10^{-4} , paired with a cosine learning rate scheduler for stable decay. We employed gradient accumulation over 4 steps (per-device batch size: 1, effective batch size: 4) and trained all models for 3 epochs. These settings provided a balance between performance and training resource constraints.

To guide the models during SFT, we prepared a diverse set of instruction prompts that encapsulated real-world legal drafting scenarios, ensuring relevance and structure. Sample prompts are shown in Table 3, and the complete set will be made public after acceptance to support reproducibility and further research in legal document generation.

7. Evaluation Metrics

To evaluate model performance in legal document generation, we adopt a multi-dimensional approach combining automatic metrics and expert judgments. This ensures coverage of surface-level similarity, semantic quality, coherence, and legal soundness.

Lexical accuracy was measured using ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005), which assess overlap between generated and reference texts. Semantic quality was assessed using BERTScore (Zhang et al., 2020) and BLANC (Vasilyev et al., 2020), which capture contextual and meaning-level alignment.

To evaluate beyond similarity, we employed G-Eval (Liu et al., 2023), a GPT-4-based framework for assessing factuality, coherence, and completeness. This allowed us to quantify performance on structured reasoning dimensions critical for legal texts.

Given the domain-specific nature of drafting, human expert evaluation was indispensable. Three legal professionals assessed each generated draft as a whole, rating them on two criteria: (i) Factual Accuracy- adherence to given instructions and legal facts, and (ii) Completeness & Comprehensiveness- coverage of all necessary legal elements. Since the target audience is ultimately legal experts, this step ensured that evaluations reflected real-world drafting needs.

To confirm reliability, we computed inter-annotator agreement (IAA) across expert ratings using Fleiss' κ (Fleiss, 1971), Cohen's κ (Cohen, 1960), Intraclass Correlation Coefficient (ICC) (Shrout and Fleiss, 1979), Krippendorff's α (Krippendorff, 2018), and Pearson correlation (Benesty et al., 2009). Higher agreement was observed for structured wrapper outputs, showing that section-wise generation aids consistent expert judgments.

Instructions:
You are an expert in legal text evaluation. You will be given:
A document description that specifies the intended content of a generated legal document.
An actual legal document that serves as the reference. A generated legal document that needs to be evaluated. Your task is to assess how well the generated document aligns with the given description while using the actual document as a reference for correctness.

Evaluation Criteria (Unified Score: 1-10)
Your evaluation should be based on the following factors:
Factual Accuracy (50%) – Does the generated document correctly represent the key legal facts, reasoning, and outcomes from the original document, as expected from the description?
Completeness & Coverage (30%) – Does it include all crucial legal arguments, case details, and necessary context that the description implies?
Clarity & Coherence (20%) – Is the document well-structured, logically presented, and legally sound?

Scoring Scale:
1-3 → Highly inaccurate, major omissions or distortions, poorly structured.
4-6 → Somewhat accurate but incomplete, missing key legal reasoning or context.
7-9 → Mostly accurate, well-structured, with minor omissions or inconsistencies.
10 → Fully aligned with the description, factually accurate, complete, and coherent.

Input Format:
Document Description:
{{doc_des}}

Original Legal Document (Reference):
{{Actual_Document}}

Generated Legal Document (To Be Evaluated):
{{Generated_Document}}

Output Format:
Strictly provide only a single integer score (1-10) as the response, with no explanations, comments, or additional text.

Table 4: The prompt is utilized to obtain scores from the G-Eval automatic evaluation methodology. We employed the GPT-4o-mini model to evaluate the quality of the generated text based on the provided prompt/input description, alongside the actual document as a reference.

8. Results and Analysis

This section presents the evaluation results of various models for legal document generation. The models were assessed using lexical-based, semantic similarity-based, automatic LLM-based, and expert evaluation metrics, as detailed in Table 5. Our findings highlight key challenges, the impact of supervised fine-tuning (SFT), and the effectiveness of the model-agnostic wrapper.

8.1. Comparative Model Performance

Among the open-source models, Qwen3-14B, LLaMA-3.1-8B-Instruct, and Gemma-3-12B-It exhibited limited performance across both lexical and semantic evaluations. More importantly, direct supervised fine-tuning (SFT) on these models led to further performance degradation. This result, while initially surprising, can be explained by several underlying factors that we analyzed in depth. One major factor was dataset diversity and imbalance. Even though the VidhikDastaavej dataset was expanded to more than 11,000 documents across 133 categories, certain specialized legal document types remained underrepresented. As a result, SFT models tended to overfit to the dominant patterns in frequently occurring categories while struggling to generalize to underrepresented ones. This behavior was evident in outputs where SFT models reproduced overly generic clauses or omitted critical sections in less frequent document types.

Another factor was the mismatch between the

instruction style used in SFT training and the structured prompting strategy employed in our wrapper. SFT relied on flat, single-shot instructions paired with full document targets, which constrained the model to a rigid learning signal. In contrast, the wrapper decomposed document generation into a two-phase workflow: planning section titles and generating section-wise content with retrieval augmentation. This structured and iterative approach aligns more closely with how legal professionals draft documents, ensuring global coherence and better factual consistency.

A third important consideration is the role of retrieval in grounding. While SFT attempted to internalize drafting knowledge directly from limited training samples, the wrapper dynamically retrieved relevant context for each section at generation time. This retrieval-augmented process reduced hallucinations, ensured clause-specific accuracy, and allowed even smaller base models to produce outputs that were coherent and factually grounded. In practice, this adaptability compensated for data sparsity and improved performance across evaluation metrics.

As shown in Table 5, wrapper-enhanced models consistently outperformed both base and SFT variants. For instance, Gemma-3-12B-It achieved an expert factual accuracy score of only 1.00 after SFT, whereas the wrapper applied over the same model yielded a score of 8.82. Similarly, LLaMA-3.1-8B-Instruct showed a marked jump in completeness when used with the wrapper. These results demonstrate that retrieval-based, modular prompting is more effective than conventional fine-tuning for long-form, high-precision domains such as legal drafting.

Finally, a closer inspection of failure cases highlighted common shortcomings of SFT models. For example, in generating a Shareholders' Agreement, SFT outputs frequently omitted key provisions such as "Governing Law" or "Termination Clauses," or invented irrelevant citations. Wrapper-enhanced models, by contrast, generated structurally complete drafts with higher factual fidelity, aided by their explicit planning and retrieval steps. Illustrative hallucination examples are included in Table 6.

Overall, these findings suggest that while SFT remains a valuable adaptation technique, its effectiveness is limited in domains like law where diversity, structure, and factual precision are paramount. The model-agnostic wrapper provides a more reliable and scalable solution, ensuring consistency across models and enabling practical deployment in real-world legal drafting tasks.

8.2. Effectiveness of MAW

One of the most promising findings of our study is the effectiveness of the model-agnostic wrap-

Models	Lexical Based Evaluation			Semantic Evaluation		Automatic LLM	Average Expert Scores	
	RL	BLEU	METEOR	BERTScore	BLANC	G-Eval	Factual Accuracy	Completeness & Compre.
Qwen3-14B	0.09	0.00	0.07	0.73	0.01	3.56	1.00	1.00
LLaMA-3.1-8B-Instruct	0.09	0.01	0.11	0.78	0.04	1.57	1.00	1.10
LLaMA-3.1-8B-Instruct SFT	0.08	0.00	0.05	0.74	0.01	1.12	1.00	1.00
Wrapper (Over LLaMA-3.1-8B)	0.15	0.04	0.18	0.79	0.19	5.15	3.30	2.20
Gemma-3-12B-It	0.09	0.01	0.10	0.76	0.02	1.13	1.00	1.00
Gemma-3-12B-It SFT	0.11	0.01	0.10	0.78	0.04	1.37	1.00	1.00
Wrapper (Over Gemma-3-12B)	0.15	0.06	0.24	0.80	0.17	6.56	8.82	7.82
GPT-4o	0.14	0.03	0.12	0.81	0.24	6.68	8.80	5.40

Table 5: Evaluation metrics for new models. LLaMA-3.1-8B-Instruct and Gemma-3-12B denote the instruction-tuned variants of their respective base models. The best scores are highlighted in bold.

Prompt	Reference Output (Correct Draft)	Generated Output (Hallucinated)
Mr. [PERSON], an elder brother, wants to authorize his brother Mr. [PERSON] by giving power of attorney to appear in the court of law for court proceedings in the matter of disputed joint property. Draft a power of attorney.	Power of Attorney I, Mr. [PERSON], hereby appoint my brother, Mr. [PERSON], to act on my behalf in all legal proceedings concerning the disputed joint property contested by our relatives. He shall represent me in court, file applications, respond to notices, and bear necessary expenses. IN WITNESS WHEREOF, I sign this on [DATE].	General Power of Attorney I, Mr. [PERSON], appoint Mr. [PERSON] to manage all my financial and real estate matters, including buying, selling, and mortgaging properties across India. He may also represent me in taxation and banking disputes. IN WITNESS WHEREOF, I grant him full authority without limitation.
Draft a Lease Agreement between Mr. [PERSON] (landlord) and Ms. [PERSON] (tenant) for a residential flat in [LOCATION] for 11 months at a monthly rent of 15,000 INR.	Lease Agreement This agreement is made between Mr. [PERSON] (landlord) and Ms. [PERSON] (tenant) for a residential flat at [LOCATION]. The term of lease shall be 11 months commencing on [DATE], with a monthly rent of 15,000 INR. The tenant agrees not to sublet the premises. Governing law: Indian Contract Act.	Lease Agreement This agreement is made between Mr. [PERSON] (landlord) and Ms. [PERSON] (tenant) for commercial office space at [LOCATION]. The term of lease shall be 5 years, with a monthly rent of 25,000 INR. The tenant may sublet the premises with prior written notice. Governing law: English Law.

Table 6: Examples of hallucinations in AI-generated legal drafting. Unlike incoherent outputs, hallucinations manifest as factual inconsistencies (e.g., inventing clauses, altering contract type, or changing law).

per in generating structured, large, and coherent legal documents. The wrapper enhances consistency across sections, ensuring logical flow and improving document quality. This method proves particularly effective for maintaining coherence in complex legal texts, overcoming the limitations of individual models. Notably, the wrapper’s outputs achieved comparable scores to GPT-4o, despite being generated using open-source models. Expert evaluations further confirm that the generated documents from wrapper-assisted models were coherent, well-structured, and legally valid, demonstrating the utility of this approach.

An additional advantage of the wrapper function is its ability to reduce hallucinations in legal text generation. Hallucinations, where the model generates factually incorrect or legally inconsistent information, pose a significant challenge in AI-generated legal documents. By enforcing a structured, step-wise document generation approach, the wrapper minimizes it by ensuring that the generated content remains grounded in the given instructions and previously generated sections.

8.3. Expert Evaluation: Factual Accuracy and Completeness

Expert evaluation provides the most reliable measure of an AI-generated document’s real-world applicability. Our findings show that factual accuracy

and completeness scores correlate strongly with expert assessments, highlighting their importance as legal-specific evaluation metrics. Models that underwent SFT struggled with maintaining factual consistency, likely due to the limited amount of the fine-tuning dataset. On the other hand, the MAW significantly improved both factual accuracy and completeness, reinforcing its role in enhancing document consistency and legal validity. Wrapper-enhanced models received high marks, with the Gemma-based wrapper achieving expert ratings of 8.82 (factual) and 7.82 (completeness), ahead of GPT-4o. This suggests wrapper-based prompting can offer performance comparable to proprietary models in specialized domains like legal NLP.

8.4. IAA Findings and Observations

Tables 7 and 8 summarize the IAA results for the factual accuracy and completeness scores, respectively. We observed moderate agreement for baseline models such as Qwen3-14B and LLaMA-3.1-8B-Instruct. However, the wrapper-enhanced configurations exhibited consistently higher agreement scores, indicating their outputs were easier for experts to evaluate consistently.

Wrapper-based variants achieved Fleiss’ κ and Krippendorff’s α above 0.80, and ICC values approaching or exceeding 0.90 for factual accuracy, highlighting strong consensus among raters. Com-

Models	Fleiss' κ	Cohen's κ	ICC	Kripp. α	Pearson Corr.
Owen3-14B	0.42	0.44	0.49	0.45	0.43
Llama-3.1-8B-Instruct	0.38	0.40	0.42	0.41	0.39
Llama-3.1-8B-Instruct SFT	0.35	0.39	0.41	0.39	0.37
Wrapper Over (Llama-3.1-8B)	0.79	0.75	0.88	0.86	0.89
Gemma-3-12b-it	0.33	0.38	0.39	0.38	0.35
Gemma-3-12b-it SFT	0.30	0.36	0.35	0.36	0.33
Wrapper Over (Gemma-3-12B)	0.81	0.80	0.91	0.90	0.92
GPT4o	0.82	0.84	0.89	0.87	0.91

Table 7: Inter-Annotator Agreement (IAA) Metrics for Factual Accuracy, evaluating consistency among expert reviewers across different models.

Models	Fleiss' κ	Cohen's κ	ICC	Kripp. α	Pearson Corr.
Owen3-14B	0.40	0.42	0.48	0.44	0.42
Llama-3.1-8B-Instruct	0.37	0.39	0.41	0.39	0.38
Llama-3.1-8B-Instruct SFT	0.36	0.38	0.40	0.38	0.36
Wrapper Over (Llama-3.1-8B)	0.73	0.70	0.85	0.83	0.87
Gemma-3-12b-it	0.34	0.37	0.37	0.36	0.34
Gemma-3-12b-it SFT	0.32	0.35	0.33	0.34	0.31
Wrapper Over (Gemma-3-12B)	0.77	0.75	0.87	0.86	0.89
GPT4o	0.78	0.79	0.88	0.86	0.90

Table 8: IAA Metrics for Completeness & Comprehensiveness, evaluating consistency among expert reviewers across different models.

pleteness scores showed similar trends, reinforcing that structured generation enhances clarity and assessment consistency. GPT-4o also demonstrated high agreement, but the best-performing wrapper-based open-source models were competitive, validating their utility as viable alternatives.

8.5. Insights from Legal Experts

In addition to numerical scoring, legal experts provided detailed qualitative feedback. Key insights:

- *Improved Structure and Coherence*: Experts appreciated that wrapper-based outputs exhibited logical progression and better adherence to legal formatting norms, particularly in section-wise organization.
- *Reduced Hallucinations*: Experts found outputs from wrapper-based models to be more factually grounded, supported by improved use of domain-specific terminology and reduced irrelevancy.
- *Linguistic Clarity and Formalism*: Continued pre-training was noted to improve the quality of formal legal language. Experts preferred drafts mimicking Indian legal writing conventions.
- *Areas for Improvement*: Minor verbosity and occasional factual inconsistencies were observed in longer drafts. Experts recommended integrating case precedents and statutory references for more robust legal drafting.

9. Ablation Study

9.1. Ablation: Impact of Retrieval Module

To quantify the contribution of the retrieval module in the Model-Agnostic Wrapper (MAW), we conducted an ablation where retrieval was disabled while keeping the rest of the structured generation pipeline intact. Removing retrieval resulted in a

decline in both expert-assessed factual accuracy (-2.2 points) and completeness (-1.7 points). Lexical metrics and semantic metrics also dropped consistently. This indicates that retrieval plays a crucial role in grounding the generation process with relevant precedents, improving both legal accuracy and contextual alignment.

9.2. Component-wise Ablation of the Wrapper

To disentangle the impact of individual components in the wrapper, we evaluated different configurations:

- *Long Prompt Only (no structure or retrieval)*: Minor improvements in lexical overlap but no noticeable change in factual accuracy.
- *Retrieval Only (flat prompt without structure)*: Showed moderate gains in completeness but lacked logical flow and legal coherence.
- *Structured Generation Only (no retrieval)*: Provided better document organization but failed to anchor content in precedent-specific context.

Only the full wrapper, combining structured generation and retrieval, consistently achieved high scores across all metrics, most notably $+4.5$ in factual accuracy and $+3.8$ in completeness (expert Likert scores). These findings confirm that both structured planning and contextual grounding are essential to improving legal document generation.

10. Conclusion and Future Work

This work introduces VidhikDastaavej, a novel large-scale dataset of private Indian legal documents, and a Model-Agnostic Wrapper (MAW) that enables structured, section-wise legal drafting across models. Our results show that while supervised fine-tuning often degraded performance, the wrapper approach consistently improved coherence, factual accuracy, and completeness, even surpassing strong proprietary models in expert evaluations. To ensure practical use, we also developed a Human-in-the-Loop system that allows experts to refine and validate drafts interactively. Together, the dataset, wrapper methodology, and expert-based evaluation establish a foundation for robust AI-assisted legal drafting in resource-constrained and evolving model settings.

Future work will expand the dataset with more documents per category, adopt rubric-style domain-specific evaluations to assess legal soundness more rigorously, and integrate factual verification modules and efficiency optimizations. These steps will enhance the scalability, reliability, and real-world usability of AI-driven legal drafting systems.

11. Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers for their constructive feedback and valuable suggestions, which greatly improved the quality of this paper. We also thank the student research assistants and legal experts from various law colleges for their dedicated efforts in annotation and expert evaluation. Their contributions were instrumental in ensuring the quality and reliability of the dataset and evaluation process. We gratefully acknowledge the support of BharatGen, India, for providing access to computational resources and hardware infrastructure used in this research. Their assistance played a vital role in model training and large-scale experimentation. The majority of this work was conducted while the first author was affiliated with the Indian Institute of Technology Kanpur. The author is currently affiliated with the University of Birmingham Dubai.

12. Limitations

Despite the promising results, some limitations remain that define directions for improvement.

First, while the `VidhikDastaavej` dataset includes over 11,000 documents across 133 categories, the test set includes only one document per category (133 total). This breadth ensures coverage across diverse document types but risks higher variance and possible sensitivity to idiosyncratic cases. To mitigate this, we relied on three independent legal experts and strong inter-annotator agreement metrics (Fleiss' κ , ICC, Krippendorff's α), ensuring reliability beyond raw averages. Nonetheless, expanding test coverage (e.g., ≥ 10 documents per category) will allow for significance testing and deeper per-type analyses.

Second, although the wrapper reduces hallucinations and factual errors, occasional inconsistencies remain. Automated evaluation methods such as G-Eval provide useful signals but can appear ad hoc. A more dedicated evaluation framework—for example, a checklist-style rubric tailored to each document type (mandatory clauses, cross-reference consistency, signature blocks, governing law, etc.), would better capture legal soundness and pinpoint error modes. This remains a key area for future refinement.

Third, the wrapper introduces moderate computational overhead. Compared to direct prompting, we observed roughly a 1.4–1.6 \times increase in inference time, with retrieval adding 60–80 ms per query. While these costs are justified by substantial quality gains, further optimizations (e.g., caching, adaptive retrieval, and prompt compression) will improve efficiency, particularly in real-world deployments where latency matters.

Finally, although evaluations involved experienced legal experts, broader deployment in professional workflows has not yet been tested. Real-world usage studies will be necessary to validate usability, trust, and adaptability across jurisdictions.

By expanding test coverage, refining domain-specific evaluation protocols, and optimizing computational efficiency, future iterations of this work can deliver even stronger, legally sound, and practically deployable AI solutions for legal drafting.

13. Ethics Statement

This research studies AI-assisted generation of legal documents, a high-stakes setting with risks related to privacy/confidentiality, bias, transparency, accountability, and potential misuse. Given the sensitive nature of legal documents, we prioritized data privacy and security in every phase of this study.

Data governance and permissions. The dataset `VidhikDastaavej` was curated in collaboration with a legal firm and used with appropriate permissions for research purposes, and no confidentiality agreements were violated during data collection and use.

Privacy and confidentiality. We de-identified documents prior to any processing (technical details in Section 4.3). Since de-identification may not fully eliminate re-identification risk in all circumstances, we treat the dataset as sensitive and limit access and use accordingly.

Bias and fairness. AI models, may inherit biases from historical legal texts, potentially affecting fairness in document generation. To reduce the risk of generating harmful or misleading content, we include expert evaluation criteria focused on factual accuracy and legal completeness, and we position model outputs as drafts requiring professional review.

Transparency, accountability, and human oversight. Transparency is crucial in legal AI applications. To improve the reliability of generated documents, we developed the Model-Agnostic Wrapper (MAW), which enforces structured text generation while minimizing hallucinations. However, AI-generated legal drafts are not substitutes for human expertise. The system is designed as an assistive tool, with a Human-in-the-Loop (HITL) mechanism that ensures legal professionals oversee and refine the generated drafts before any official use.

14. Bibliographical References

- Satanjeev Banerjee and Alon Lavie. 2005. [ME-TEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.
- Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. In *Legal Knowledge and Information Systems*, pages 3–12. IOS Press.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.
- Ashish Chouhan and Michael Gertz. 2024. [Lex-Drafter: Terminology drafting for legislative documents using retrieval augmented generation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10448–10458, Torino, Italia. ELRA and ICCL.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. [CLERC: A dataset for U. S. legal case retrieval and retrieval-augmented analysis generation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7898–7913, Albuquerque, New Mexico. Association for Computational Linguistics.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Xiaolin Li, Lei Huang, Yifan Zhou, and Changcheng Shao. 2021. Tst-gan: A legal document generation model based on text style transfer. In *2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE)*, pages 90–93. IEEE.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Kumar Nigam, Arnab Bhattacharya, and Ashutosh Modi. 2022. [Semantic segmentation of legal documents via rhetorical roles](#). In *Proceedings of the Natural Language Processing Workshop 2022*, pages 153–171, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- 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 Language Processing Workshop 2023*, pages 73–84.
- Gianluca Moro, Nicola Piscaglia, Luca Ragazzi, and Paolo Italiani. 2024. Multi-language transfer learning for low-resource legal case summarization. *Artificial Intelligence and Law*, 32(4):1111–1139.

- Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11085–11093.
- Shubham Nigam, Anurag Sharma, Danush Khanna, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2024. [Legal judgment reimaged: PredEx and the rise of intelligent AI interpretation in Indian courts](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4296–4315, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Shubham Kumar Nigam, Deepak Patnaik Balaramamahanthi, Shivam Mishra, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. [NYAYAANUMANA and INLEGALLAMA: The largest Indian legal judgment prediction dataset and specialized language model for enhanced decision analysis](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11135–11160, Abu Dhabi, UAE. Association for Computational Linguistics.
- Shubham Kumar Nigam, Navansh Goel, and Arnab Bhattacharya. 2022. [nigam@coliee-22: Legal case retrieval and entailment using cascading of lexical and semantic-based models](#). In *JSAI International Symposium on Artificial Intelligence*, pages 96–108. Springer.
- Shubham Kumar Nigam, Shubham Kumar Mishra, Ayush Kumar Mishra, Noel Shallum, and Arnab Bhattacharya. 2023. Legal question-answering in the indian context: Efficacy, challenges, and potential of modern ai models. *arXiv preprint arXiv:2309.14735*.
- Vasile Păiș, Maria Mitrofan, Carol-Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the romanian legal domain. In *Proceedings of the natural legal language processing workshop 2021*, pages 9–18.
- Rithik Raj Pandey, Sarthak Khandelwal, Satyam Srivastava, Yash Triyar, and Muquitha Almas. 2024. [LEGALSEVA - AI-powered legal documentation assistant](#). *International Research Journal of Modernization in Engineering, Technology and Science*, 6(3):6418–6423.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Atharv Patil, Kartik Bapna, and Ayush Shah. 2024. [Legal docgen using ai: Your smart doc generator](#). *International Journal of Novel Research and Development*, 9(5):536–543.
- Luca Ragazzi, Gianluca Moro, Stefano Guidi, and Giacomo Frisoni. 2024. Lawsuit: a large expert-written summarization dataset of italian constitutional court verdicts. *Artificial Intelligence and Law*, pages 1–37.
- TYSS Santosh, Apolline Isaia, Shiyu Hong, and Matthias Grabmair. 2024. Hiculr: Hierarchical curriculum learning for rhetorical role labeling of legal documents. *arXiv preprint arXiv:2409.18647*.
- Patrick E Shrouf and Joseph L Fleiss. 1979. Intra-class correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.
- 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. *arXiv preprint arXiv:2210.07544*.
- Takaaki Tateishi, Sachiko Yoshihama, Naoto Sato, and Shin Saito. 2019. Automatic smart contract generation using controlled natural language and template. *IBM Journal of Research and Development*, 63(2/3):6–1.
- Chroma Team. 2023. [Chroma: Open-source embedding database](#). Accessed: July 2025.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Yu Tong, Weiming Tan, Jingzhi Guo, Bingqing Shen, Peng Qin, and Shuaihe Zhuo. 2022. Smart contract generation assisted by ai-based word segmentation. *Applied Sciences*, 12(9):4773.
- Santosh T.y.s.s, Mahmoud Aly, Oana Ichim, and Matthias Grabmair. 2025a. [LexGenie: Automated generation of structured reports for European court of human rights case law](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 672–683, Vienna, Austria. Association for Computational Linguistics.
- Santosh T.y.s.s, Youssef Farag, and Matthias Grabmair. 2025b. [CoPERLex: Content planning with event-based representations for legal case summarization](#). In *Findings of the Association for*

Computational Linguistics: NAACL 2025, pages 1016–1032, Albuquerque, New Mexico. Association for Computational Linguistics.

Santosh T.y.s.s and Elvin Quero Hernandez. 2025. [LexKeyPlan: Planning with keyphrases and retrieval augmentation for legal text generation: A case study on European court of human rights cases](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–436, Vienna, Austria. Association for Computational Linguistics.

Oleg V. Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. [Fill in the BLANC: human-free quality estimation of document summaries](#). *CoRR*, abs/2002.09836.

Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Nigam, Shouvik Guha, Koustav Rudra, and Kripabandhu Ghosh. 2023. Lims—the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on indian court cases. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12451–12474.

Haifeng Wei. 2024. Intelligent legal document generation system and method based on knowledge graph. In *Proceedings of the 2024 International Conference on Machine Intelligence and Digital Applications*, pages 350–354.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.