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

Anonymous ACL submission

Abstract

Automating legal document drafting can enhance efficiency, reduce manual workload, and streamline legal workflows. However, the structured generation of private legal documents remains underexplored, particularly in the Indian legal context due to limited public data and model adaptation challenges. We propose a Model-Agnostic Wrapper (MAW), a flexible, two-stage generation framework that first produces section titles and then generates section-wise content using retrieval-based prompts. This wrapper decouples generation from any specific model, enabling compatibility with a range of open and closed-source LLMs, and ensuring coherence, factual alignment, and reduced hallucination. To enable practical use, we build a Human-in-the-Loop Document Generation System, an interactive interface where users can input document types, refine sections, and iteratively generate structured drafts. The tool supports real-world legal workflows and will be made publicly accessible upon acceptance with privacy and security safeguards. Comprehensive evaluations, including expert-based assessments, demonstrate that the wrapper-based approach substantially improves document quality over baseline and fine-tuned models. Our framework establishes a scalable and adaptable path toward structured AI-assisted legal drafting in the Indian domain.

1 Introduction

Automating legal document generation can significantly improve efficiency and accessibility in legal workflows. While LLMs have been widely used for legal tasks such as judgment prediction, case summarization, and retrieval, their application to private legal document generation 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.

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, model codes, and user interface accessible via an anonymous repository[1]. After acceptance, we will release the tool publicly with privacy, security, and copyright considerations to facilitate general use.

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 seen significant progress in the legal domain, with applications spanning 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 Named Entity Recognition (NER) (Păiș et al., 2021). These advances enable more accurate, explainable, and efficient legal decision support systems. In the Indian context, much of the prior research has focused on public legal judgments, particularly from the Supreme Court and High Courts, to support tasks like retrieval, reasoning, and explainability (Chalkidis et al., 2020).

Several benchmark datasets have emerged to facilitate Legal AI research in India. The Indian Legal Documents Corpus (ILDC) (Malik et al., 2021) and PredEx (Nigam et al., 2024) support judgment prediction and rationale extraction. Work on rhetorical role labeling (Bhattacharya et al., 2019; Malik et al., 2022) and factual segmentation (Nejadgholi et al., 2017) has further enriched structural understanding. HiCuLR (Santosh et al., 2024) introduced hierarchical curriculum learning for segmenting Indian legal judgments, while recent studies have also addressed legal NER using large-scale pretrained models (Vats et al., 2023).

Beyond judgment-focused tasks, legal document generation has received increasing attention globally. Early work includes rule-based and controlled natural language drafting (Tateishi et al., 2019), AI-assisted segmentation (Tong et al., 2022), and text style transfer models (Li et al., 2021). Knowledge graph-based methods have been applied to ensure coherence and semantic fidelity in legal generation tasks (Wei, 2024). 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.

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 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

[1] https://anonymous.4open.science/r/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.

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, curated a 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.

4.1 Dataset Composition and Diversity

The dataset encompasses a wide variety of License Agreements, Severance Agreements, Stock Option Agreements, Consulting Agreements, Asset Purchase Agreements, and more. By incorporating multiple document types, VidhikDastaavej captures the 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 reserved for testing. The dataset covers 133 legal document categories in the training set and 133 categories in the test set, offering a broad representation of real-world legal drafts.

To ensure balanced exposure to different legal drafting styles, we structured the dataset to include a well-distributed mix of document types. A visual distribution of the top 15 most frequent document categories is presented in Appendix Figure 7. The detailed document type distributions are available in the anonymous GitHub repository and the uploaded dataset folder. This diversity is critical for training models that generalize across different legal document formats, improving their usability in real-world legal drafting.

4.2 Data Anonymization and Ethical Considerations

To comply with privacy regulations and ethical standards, all documents in VidhikDastaavej underwent a rigorous anonymization process. We employed Spacy Named Entity Recognition (NER) tools to systematically replace personal identifiers, such as names, addresses, and confidential details, with placeholders. This preserves document integrity while ensuring that no personally identifiable information (PII) is exposed, making the dataset safe for research and model development. A sample document showing after anonymization in the Appendix Table 8.

4.3 Significance of the Dataset

Unlike previous datasets that primarily focus on court judgments or a single category of legal texts, VidhikDastaavej provides a comprehensive representation of private legal documentation in India. This enables language models to learn the intricacies of Indian legal terminology, structural conventions, and drafting practices. The dataset serves as a foundational resource for training and evaluating legal document generation models, facilitating the development of AI-powered tools capable of assisting legal practitioners in drafting structured, coherent, and legally sound documents efficiently.

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.

5.1 Two-Phase Structured Document Generation

The MAW employs a two-phase workflow (Figure 1) 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

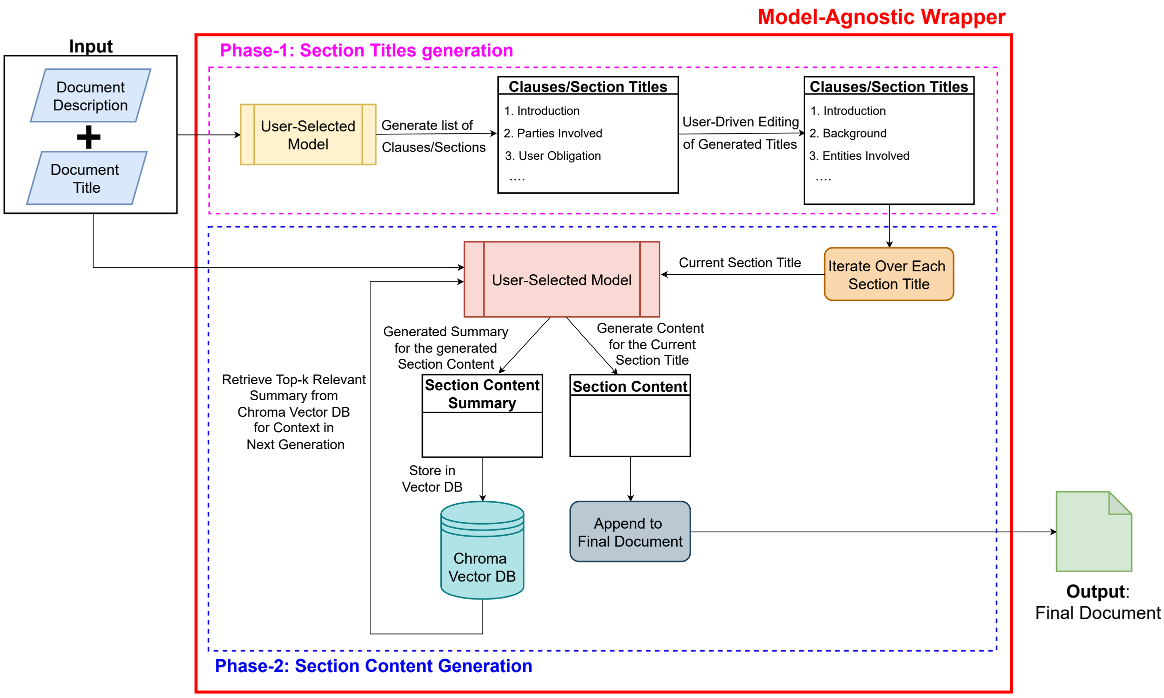

Figure 1: Wrapper flow diagram

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 the 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 separately, 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.

6 Experimental Setup

To benchmark our pipeline's performance and assess our wrapper's effectiveness, we conducted 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. Examples of these prompts and instructions are included in Appendix Table 5.

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 Appendix 5, and the complete set will be made public after acceptance to support reproducibility and further research in legal document generation.

7 Evaluation Metrics

To assess the performance of the legal document generation models, we adopt a multi-faceted evaluation approach that includes lexical-based, semantic similarity-based, automatic LLM-based, and expert evaluation metrics. Since legal document drafting requires precision, coherence, and adherence to legal norms, these evaluation methods ensure a comprehensive assessment of model performance.

1. **Lexical-based Evaluation:** We utilized standard lexical similarity metrics, including Rouge-L (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005). These metrics measure the overlap and order of words between the generated explanations and the reference texts, providing a quantitative assessment of the lexical accuracy of the model outputs.
2. **Semantic Similarity-based Evaluation:** To capture the semantic quality of the generated explanations, we employed BERTScore (Zhang et al., 2020), which measures the semantic similarity between the generated text and the reference explanations. Additionally, we used BLANC (Vasilyev et al., 2020), a metric that estimates the quality of generated text without a gold standard, to evaluate the model's ability to produce semantically meaningful and contextually relevant explanations.
3. **Automatic LLM-based Evaluation:** This evaluation is crucial for assessing structured argumentation and legal correctness. We employ G-Eval (Liu et al., 2023), a GPT-4-based framework designed for NLG assessment, which leverages chain-of-thought reasoning and structured form-filling to improve alignment with human judgment. This evaluation provides insights into coherence, factual accuracy, and completeness beyond traditional similarity metrics. The evaluation prompt used for obtaining G-Eval scores is detailed in Appendix Table 7.
4. **Expert Evaluation:** To ensure a rigorous and unbiased evaluation, we engaged legal professionals with expertise in drafting and reviewing legal documents. These experts were recruited through professional legal networks and academia. Each expert was compensated for their time and expertise at a fair market rate, ensuring that their efforts were adequately acknowledged. Given the domain-specific nature of legal documents, human expert evaluation is necessary to assess the practical utility of AI-generated texts. Unlike existing evaluation approaches that primarily rely on lexical or semantic similarity, this expert-driven evaluation ensures that AI-generated legal content meets professional standards. We introduce two key evaluation criteria in this category:
   (a) **Factual Accuracy:** This metric evaluates whether the generated document strictly ad-

heres to the given instructions, accurately represents legal facts, and avoids hallucination or misinformation. In legal drafting, factual inaccuracies can lead to severe consequences, making this metric crucial for ensuring the reliability of AI-generated legal documents.

   (b) **Completeness and Comprehensiveness:** This metric assesses how well the generated document covers all necessary legal aspects. A legally sound document should include all relevant arguments, clauses, and supporting details. Omissions or inconsistencies in legal drafting can render a document ineffective or legally invalid.

5. **Inter-Annotator Agreement (IAA) for Expert Evaluation:** To further ensure the consistency and reliability of expert evaluations, we conducted an IAA analysis across two key dimensions: *Factual Accuracy* and *Completeness & Comprehensiveness*. Three legal experts independently rated outputs without knowledge of the generating model to avoid bias.

   We employed five widely used IAA metrics:

   - *Fleiss' Kappa* (Fleiss, 1971): Measures agreement among more than two annotators over categorical data.
   - *Cohen's Kappa* (Cohen, 1960): Evaluates pairwise agreement between two annotators, accounting for chance agreement.
   - *Intraclass Correlation Coefficient (ICC)* (Shrout and Fleiss, 1979): Captures reliability when measuring continuous scores from multiple raters.
   - *Krippendorff's Alpha* (Krippendorff, 2018): A robust reliability measure applicable to various data types and missing values.
   - *Pearson Correlation Coefficient* (Benesty et al., 2009): Measures the linear correlation between pairs of raters across scores.

   These metrics provide a comprehensive view of expert consensus. Higher agreement was observed for models generating more structured outputs, particularly those using our wrapper approach, confirming that coherent, section-wise generation aids in consistent human evaluation.

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 2. 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 showed limited performance across lexical and semantic evaluations. Interestingly, direct supervised fine-tuning (SFT) on these models led to further degradation, a finding that warrants deeper analysis. We attribute this decline not merely to dataset limitations, but also to three key factors:

First, we observed indications of potential overfitting, especially for SFT models evaluated on out-of-domain or underrepresented document types. Since the fine-tuning dataset, though extended, remained relatively narrow in terms of legal diversity and structural variation, the models likely learned to memorize patterns rather than generalize drafting logic.

Second, the instruction format mismatch between the SFT and wrapper strategies played a pivotal role. While SFT trains models on full-target outputs conditioned on flat instructions, our wrapper follows a hierarchical and retrieval-augmented prompting approach that decomposes the task into smaller, interpretable sections. This structured generation process aligns better with legal drafting norms and mimics how human lawyers approach document composition.

Third, the wrapper approach dynamically integrates relevant information at generation time via retrieval, allowing contextually grounded and clause-specific responses. This adaptability compensates for data sparsity and enhances factual consistency, contributing to superior performance even with smaller base models.

As shown in Table 2, wrapper-enhanced models significantly outperform both SFT and base models across all evaluation dimensions. For instance, Gemma-3-12B-It + Wrapper achieves an expert factual accuracy score of 8.82, compared to 1.00 for the same model fine-tuned via SFT. This underscores that retrieval-augmented and modular prompting strategies can be more effective than conventional SFT in low-resource, high-precision domains like legal drafting.

Some examples illustrating hallucinations in SFT outputs are included in Appendix Table 6,

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 2: 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 for each metric are highlighted in bold.

highlighting common failure cases such as omitted clauses or invented citations. These reinforce the need for modular, grounded generation strategies over monolithic fine-tuning approaches for complex legal tasks.

8.2 Effectiveness of Model-Agnostic Wrapper

One of the most promising findings of our study is the effectiveness of the model-agnostic wrapper 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, stepwise 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 3 and 4 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. Completeness 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

Models	Fleiss' κ	Cohen's κ	ICC	Kripp. α	Pearson Corr.
Qwen3-14B	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 3: 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.
Qwen3-14B	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 4: Inter-Annotator Agreement (IAA) Metrics for Completeness & Comprehensiveness, evaluating consistency among expert reviewers across different models.

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 irrelevant content.
- *Linguistic Clarity and Formalism*: Continued pretraining 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 (e.g., BLEU) and semantic metrics (e.g., BERTScore) 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 study presents a structured and model-independent approach to legal document generation in the Indian context. We introduce a Model-Agnostic Wrapper, a two-stage framework that enhances long-form legal drafting by decoupling content generation from specific LLM architectures. The wrapper first generates section titles and then produces section-wise content, integrating retrieval-based context to ensure coherence, consistency, and factual accuracy. Our findings demonstrate that while standard fine-tuning on limited datasets does not always lead to improvements, the wrapper-based approach significantly improves performance across both automatic and expert-based evaluation metrics. This confirms the potential of structured generation strategies that are agnostic to the underlying language model, making the approach robust, scalable, and compatible with evolving LLMs.

Future work will focus on further expanding and diversifying the dataset to include additional categories of legal documents and increasing the number of expert-labeled samples. We also plan to integrate RLHF, and advanced factual verification modules to further improve factual consistency and reduce hallucinations in AI-generated legal drafts. This research lays a foundation for the development of adaptable, resource-efficient, and legally sound AI systems to support legal professionals in structured document drafting.

Limitations

Despite the advancements presented in this work, several limitations remain that highlight important opportunities for future research and system improvement.

First, while the expanded `VidhikDastaavej` dataset includes over 11,000 documents across 133 categories, an imbalance persists in the distribution of document types. Certain legal formats, such as highly specialized contracts or region-specific affidavits, are underrepresented. This may affect model generalization and robustness. Ongoing efforts are needed to incorporate more diverse, representative, and jurisdictionally varied examples of legal drafts to broaden the dataset's applicability.

Second, although our Model-Agnostic Wrapper (MAW) significantly enhances factual accuracy and structural coherence, it does not fully eliminate hallucinations or factual inconsistencies. These issues, occasionally observed in complex legal scenarios, can arise from limitations in the underlying language models or insufficient contextual knowledge. Future work could integrate external legal knowledge bases, precedents, or fact-verification modules to further reduce hallucinations and improve factual reliability.

Third, the MAW introduces a modest computational overhead due to its two-phase generation strategy and vector-based retrieval. Empirically, we observe an approximate 1.4–1.6$\times$ increase in end-to-end inference time compared to standard prompting, with retrieval adding 60–80 ms per query. While this increase is justified by substantial quality gains, particularly in legal factuality and completeness, further optimizations (e.g., caching, adaptive retrieval, and prompt compression) will be valuable for improving efficiency and supporting deployment in real-world legal environments where latency and cost are critical.

Lastly, while our evaluation involves domain experts, the system has not yet been deployed in production legal workflows. Broader validation through user studies involving practicing legal professionals across sectors would offer deeper insights into practical utility, usability, and trustworthiness.

By addressing these challenges, future iterations of our system can be made more robust, efficient, and adaptable to the evolving needs of legal professionals across domains and jurisdictions.

Ethics Statement

This research acknowledges the ethical concerns associated with AI-driven legal document generation, particularly in privacy, bias, transparency, and accountability. Given the sensitive nature of legal documents, we prioritized data privacy and security in every phase of this study. The dataset `VidhikDastaavej` was curated in collaboration with a legal firm, ensuring strict compliance with ethical guidelines. All documents were acquired with appropriate permissions, and no confidentiality agreements were violated during data collection and use.

To safeguard privacy, we implemented a robust anonymization process. Sensitive information was systematically replaced with markers while preserving document structure and legal context. Named Entity Recognition (NER)-based redaction techniques were used to mask personal identifiers, followed by manual verification to ensure completeness and accuracy. This guarantees that no personally identifiable details remain in the dataset while maintaining its relevance for AI training.

AI models, may inherit biases from historical legal texts, potentially affecting fairness in document generation. To mitigate this, we introduced expert-based evaluation criteria focusing on factual accuracy and completeness to ensure generated documents adhere to legal standards and do not propagate biased or misleading content.

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.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: 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.

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.

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 Legal 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 Legal 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.

Isar Nejadgholi, Renaud Bougueng, and Samuel Witherspoon. 2017. A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In *Legal knowledge and information systems*, pages 125–134. IOS Press.

Shubham Nigam, Anurag Sharma, Danush Khanna, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2024. Legal judgment reimagined: 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.

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 Shrout and Joseph L Fleiss. 1979. Intraclass 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.

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. Llms–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.

A HITL Document Generation System: A User Guide

A.1 Overview

The Human-in-the-Loop (HITL) Document Generation System is a platform designed to create legal documents based on user inputs. Users specify the document type, provide section details, and generate structured legal documents tailored to their needs.

A.2 User Interface Guide

A.2.1 Entering Document Information

As shown in Figure 2, users begin by providing essential details about the document:

- **Document Type:** Enter the type of legal document (e.g., “Service Agreement”).
- **Description:** Provide additional context or details to customize content.
- **AI Model Selection:** Choose the LLM for document generation.
- **Begin Button:** Initiates section title generation.
- **Clear All Button:** Resets all input fields.

A.2.2 Managing Document Sections

- After clicking **Begin**, section names appear (e.g., “Parties,” “Terms and Termination”).
- Each section has the following controls:
  - **Modify:** Edit the section title.
  - **Delete:** Remove a section.
  - **Copy:** Copy the section title for reuse.
- **Add New Sections:** Click the green plus (+) icon to insert additional sections
- **Saving Titles:** Save section names before content generation.
- Figure 3 illustrates the process of editing section titles through the interface, while Figure 4 demonstrates how the addition of new section titles, along with the option to save the final titles, is seamlessly integrated within the interface.

A.2.3 Generating Section Content

Once the section titles have been finalized, the content generation process can commence, as illustrated in Figure 5. A high-level overview of the available options within the interface is provided below:

- **Stop Button:** Allows users to halt the content generation process if necessary.
- **Manual Editing:** Provides users the flexibility to refine and modify the generated content as required.
- **Copy Function:** Facilitates copying the generated section content for use in external applications or documents.

A.2.4 Exporting the Document

After finalizing the document, users can export it in different formats as shown in Figure 6:

- **Combine All:** Merges section titles and generated content into a complete document.
- **Combine Titles Only:** Exports only section titles.

A.3 Conclusion

The HITL Document Generation System provides an intuitive interface for users to generate and refine legal documents efficiently. With a structured workflow, AI-assisted drafting, and manual oversight, the system streamlines the creation of contracts, petitions, and other legal documents while maintaining coherence and accuracy. The integration of HITL ensures that legal professionals can leverage AI for drafting while retaining full control over the final output.

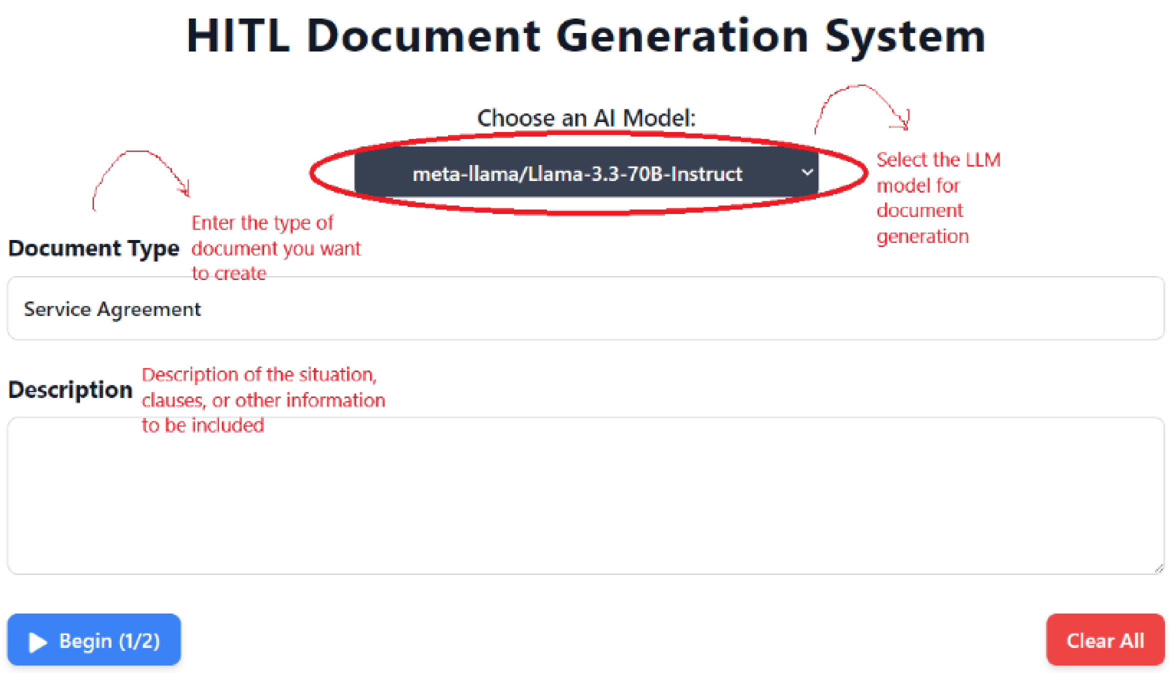

Figure 2: Document Information Entry Interface

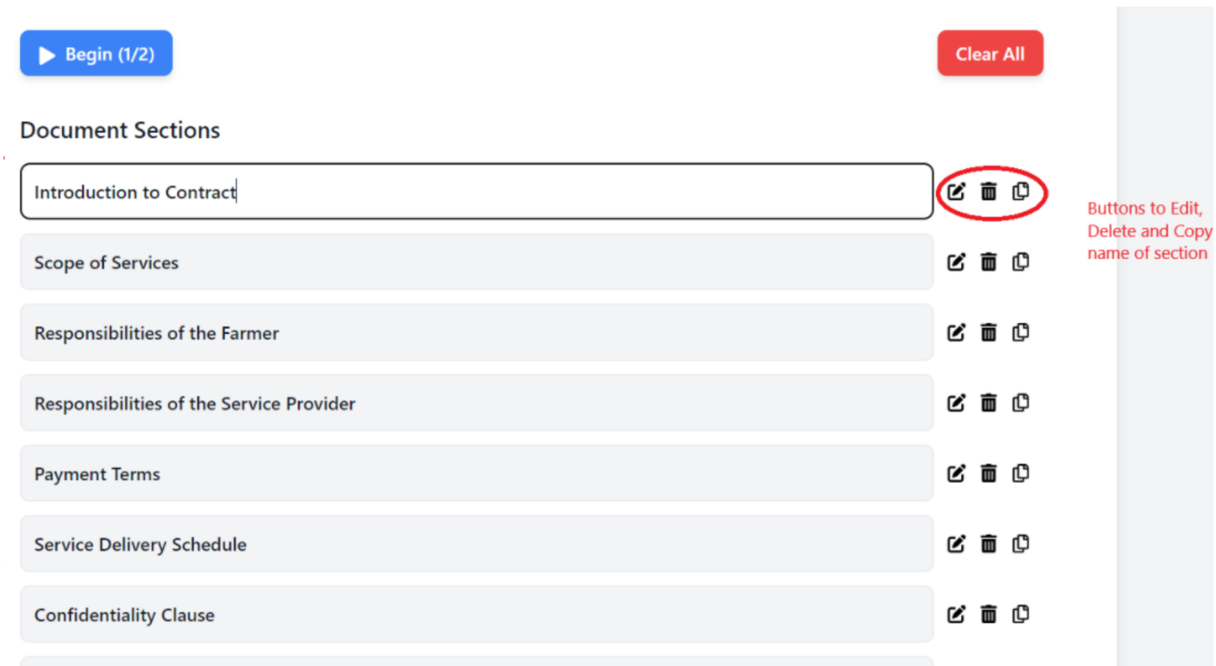

Figure 3: Editing Generated Document Sections

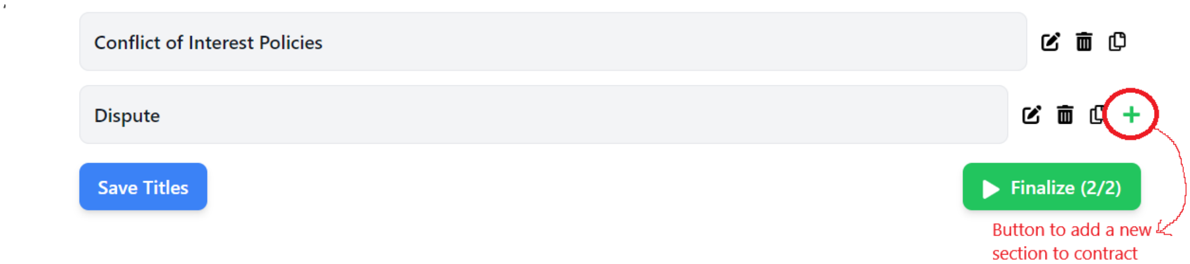

Figure 4: Adding Document Sections

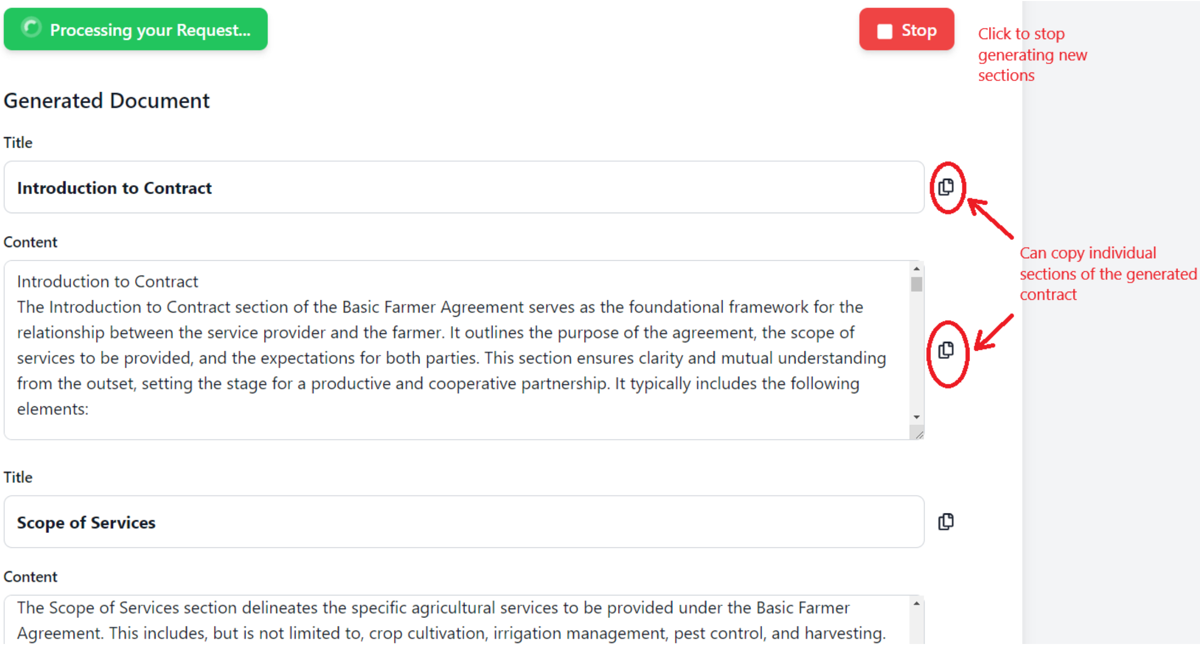

Figure 5: Generating Section Content

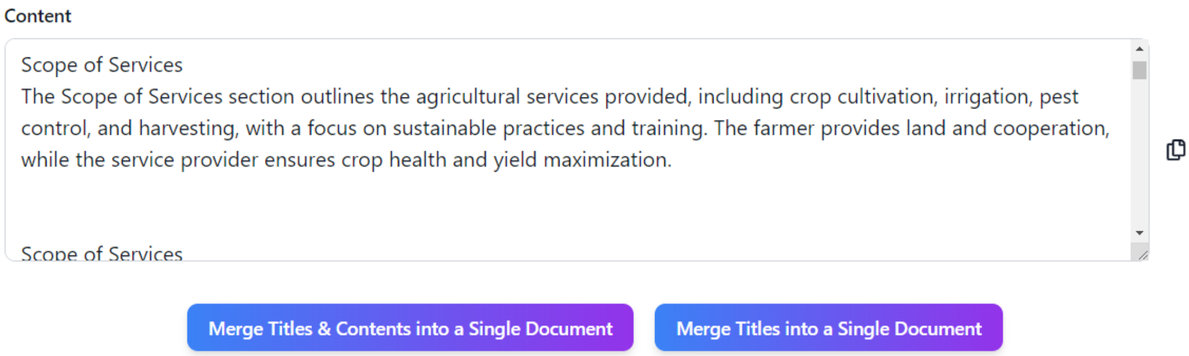

Figure 6: Exporting the Document

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 5: Categories and Corresponding Prompts for Legal Document Generation

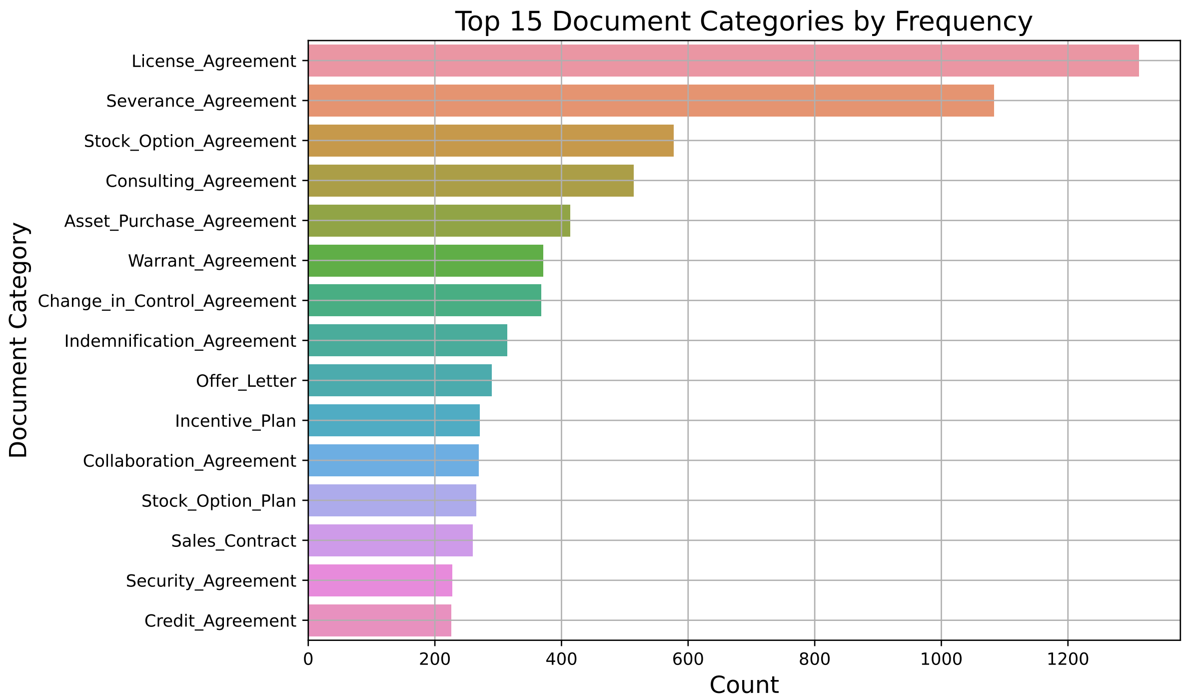

Figure 7: Top 15 Document Categories in the `VidhikDastaavej` Dataset based on Frequency.

Prompt	Actual Output	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 in a dispute contested by their other relatives. Draft a power of attorney.	**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]	socketsajs Tortildeneselildeildenild enajseselUNEomor ildeTortajsagueeselilden agogueUNEibbon attentesel ULKLEV tortomoreltasilde agateildeinasULKagenUNE ildeildeilde inflictilenamesildeildeagen LEVLEVULKildeinasLEVilNE TortildeLEVildeagogue ildeagateildenilde tortteNELEVinishedULKildeinas Tort attentLEVildenLEVLEVteg TortLEV attentLEV agate attent tort attentildeULK ULKULKagogueagateldenULK attentildeLEVULKULK tortUNEesonildeULK

Table 6: Example of hallucinations in AI-generated (LLaMA-3.1-8B-Instruct after SFT) legal document drafting. The model produced unintelligible output instead of a coherent Power of Attorney document. Non-ASCII characters have been removed to avoid compilation errors.

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 7: 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.

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