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 Large Language Models (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 valu-

able 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 made significant advancements in the legal domain, particularly in judgment prediction, case summarization, semantic segmentation, legal Named Entity Recognition (NER), and case retrieval (Chalkidis et al., 2020). In India, research efforts have primarily focused on public legal judgment cases, emphasizing explainability, retrieval, and reasoning to enhance judicial transparency and interpretability.

Several datasets have been developed to support AI applications in the Indian legal domain. The Indian Legal Documents Corpus (ILDC) (Malik et al., 2021) and PredEx (Nigam et al., 2024) provide large-scale datasets for judgment prediction with explanations. 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). Additionally, research on segmenting legal documents into distinct functional parts has been explored (Šavelka and Ashley, 2018), along with semi-supervised approaches for distinguishing factual from non-factual sentences using fastText classifiers (Nejadgholi et al., 2017). Significant progress has been made in rhetorical role labeling for Indian legal texts. Prior work has proposed models such as CRF-BiLSTM (Bhattacharya et al., 2019) and the MTL framework for the classification of legal sections (Malik et al., 2022). Recent advancements include the HiCuLR framework, which employs hierarchical curriculum learning for rhetorical role labeling (Santosh et al., 2024). Furthermore, large annotated datasets have been used in legal NER tasks, helping to extract named entities from Indian case laws (Vats et al., 2023). Several studies have explored the adaptability of large-scale pretrained models such as GPT-3.5 Turbo, LLaMA-2, and Legal-BERT for Indian legal applications (Chalkidis et al., 2020).

Despite these advancements in legal text processing and retrieval, the automation of private legal document drafting remains largely unexplored in the Indian context. While previous research has focused on processing and analyzing legal judgments, little work has been done on AI-driven generation of legal drafts. In other legal systems, various methodologies have been explored, including controlled natural language and template-based drafting (Tateishi et al., 2019), AI-assisted word segmentation for legal contracts (Tong et al., 2022), text style transfer for legal document generation (Li et al., 2021), and knowledge graph-based approaches to improve document structure and coherence (Wei, 2024).

More recently, research has begun to explore AI-powered legal document generation. TST-GAN introduced a text style transfer-based generative adversarial network for legal text generation (Li et al., 2021). Another approach leveraged knowl-

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

edge graphs to generate structured legal documents, ensuring semantic accuracy (Wei, 2024). Additionally, fine-tuned large language models have been investigated for drafting contracts and other legal documents (Lin and Cheng, 2024). AI-driven legal documentation assistants such as LEGAL-SEVA (Pandey et al., 2024) have been developed to streamline document drafting processes. The Legal DocGen Generator (Patil et al., 2024) provides a structured approach to automating legal document generation. Other studies have focused on integrating judgment prediction with legal retrieval to enhance generative models (Qin et al., 2024).

With the rise of generative AI in legal drafting, models such as Legal-BERT and LLaMA-based architectures have been fine-tuned for domain-specific text generation (Lin and Cheng, 2024). However, challenges remain due to the lack of publicly available datasets for private legal documents, which are often confidential. While some research has explored AI-powered legal assistants (Imogen[1] et al., 2024) and automated legislative drafting (Lin and Cheng, 2024), many existing models still struggle with hallucinations, inconsistencies, and domain-specific reasoning.

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 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.

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`.

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. The detailed document type distributions for the training and test sets are provided in the Anonymous Link and the uploaded dataset due to a large number of lists. 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 how the document will be after anonymization is present in the Appendix Table 6.

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

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 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) 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 the performance of `NyayaShilpi` and assess the effectiveness of our wrapper, we conducted instruction tuning on various open-source models and compared their performance against GPT-4o. Due to space constraints, complete details on hyperparameters and training configurations are provided in Appendix 7.

6.1 Instruction Tuning of Open-Source Models

We fine-tuned select open-source models while directly evaluating others without additional training. The instruction-tuned models include Phi-3 mini, which was fine-tuned using the Unsloth framework for efficiency, and LLaMA-2-7B-Chat CPT, which underwent continued pretraining (CPT) on a large

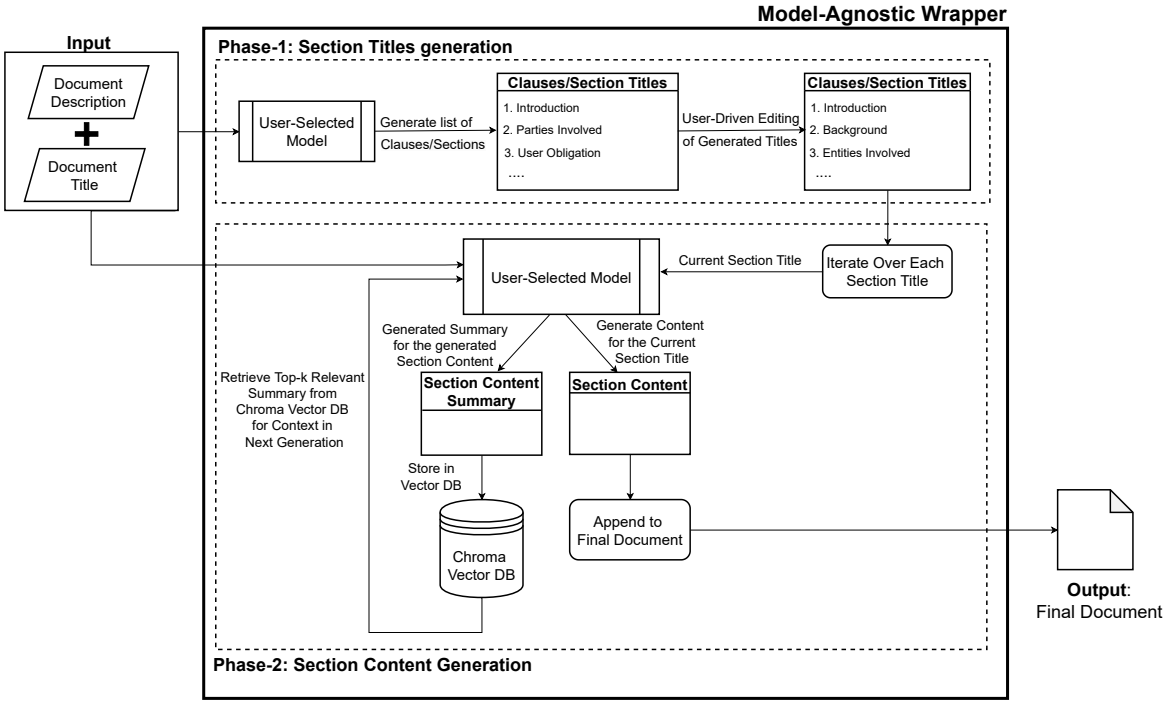

Figure 1: Wrapper flow diagram

corpus of Indian legal cases. Further fine-tuning on private legal documents led to LLaMA-2-7B-Chat CPT+SFT (`NyayaShilpi`), which serves as our primary domain-adapted model. Additionally, we fine-tuned LLaMA-3-8B-Instruct SFT to assess improvements in structured legal drafting.

In contrast, some models were directly evaluated without any fine-tuning. LLaMA-2-7B-Chat and LLaMA-3-8B-Instruct were used in their original forms as baselines to examine how well general-purpose legal models perform without additional domain-specific training. This allows us to compare whether instruction tuning meaningfully improves legal document generation quality.

For instruction tuning, we designed specialized prompts and instruction sets tailored to legal drafting. These instructions provided structured examples, ensuring the models understood the nuances of different legal document types. Examples of these prompts and instructions are included in Appendix Table 3.

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.

7 Experimental Setup and 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 perfor-

mance 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 3, and the complete set will be made public after acceptance to support reproducibility and further research in legal document generation.

8 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 scores (Rouge-1, Rouge-2, and 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 5.
4. **Expert Evaluation:** Given the domain-specific nature of legal documents, human expert evaluation is necessary to assess the practical utility of AI-generated texts. We introduce two key evaluation criteria in this category:
   (a) **Factual Accuracy:** This metric evaluates whether the generated document strictly adheres 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. 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.

   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. This process ensures that evaluations reflect real-world legal drafting practices and maintain high reliability.

9 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 continued pretraining (CPT) and supervised fine-tuning (SFT), and the effectiveness of the model-agnostic wrapper.

9.1 Comparative Model Performance

Among the Open-source models, Qwen3-14B, LLaMA-3.1-8B-Instruct, and Gemma-3-12B-It

Models	Lexical Based Evaluation					Semantic Evaluation		Automatic LLM	Average Expert Scores	
	Rouge-1	Rouge-2	Rouge-L	BLEU	METEOR	BERTScore	BLANC	G-Eval	Factual Accuracy	Completeness & Comprehensiveness
Qwen3-14B	0.1312	0.0104	0.0949	0.0003	0.0725	0.7290	0.0065	3.5551	1.0000	1.0000
LLaMA-3.1-8B-Instruct	0.1867	0.0438	0.0907	0.0095	0.1070	0.7816	0.0383	1.5741	1.0000	1.1000
LLaMA-3.1-8B-Instruct SFT	0.0887	0.0123	0.0764	0.0001	0.0467	0.7419	0.0061	1.1190	1.0000	1.0000
Wrapper (Over LLaMA-3.1-8B-Instruct)	0.3392	0.1484	0.1521	0.0437	0.1821	0.7868	0.1903	5.1540	3.3000	2.2000
Gemma-3-12B-It	0.1961	0.0560	0.0925	0.0102	0.0974	0.7619	0.0227	1.1303	1.0000	1.0000
Gemma-3-12B-It SFT	0.1775	0.0331	0.1071	0.0065	0.1002	0.7777	0.0413	1.3672	1.0000	1.0000
Wrapper (Over Gemma-3-12B-It)	**0.4103**	**0.1858**	**0.1467**	0.0620	**0.2447**	0.8029	0.1746	6.5556	8.8182	7.8182
GPT-4o	0.2486	0.1391	**0.1433**	0.0271	0.1224	**0.8095**	**0.2385**	**6.6792**	8.8000	5.4000

Table 2: Evaluation metrics for new models. LLaMA-3.1-8B-Instruct and Gemma-3-12B-It denote the instruction-tuned variants of their respective base models. The best scores for each metric are highlighted in bold.

showed limited performance in both lexical and semantic evaluation. Direct fine-tuning via SFT on these models resulted in further degradation, likely due to insufficient adaptation to complex legal drafting tasks, despite an expanded and diversified dataset.

In contrast, wrapper-enhanced models demonstrated significant improvements across all metrics. The wrapper applied over Gemma-3-12B-It, in particular, achieved the highest expert scores, outperforming even GPT-4o in factual accuracy. The wrapper's structured generation strategy provided better alignment with legal drafting norms, improved coherence, and minimized hallucination. This adequate dataset improved model generalization, especially under the wrapper-enhanced setting. Nevertheless, SFT models continued to struggle, reinforcing that model architecture and generation strategy are critical factors beyond dataset size alone. Some examples of hallucinations encountered in model outputs are provided in Appendix Table 4, due to space constraints.

9.2 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 hallucinations by ensuring that the generated content remains grounded in the given instructions and previously generated sections.

9.3 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 7.81 (completeness), slightly ahead of GPT-4o. This suggests wrapper-based prompting can offer performance comparable to proprietary models in specialized domains like legal NLP. Further analysis of expert feedback, detailed in Appendix Section **??**, provides deeper insights into how different models handle legal drafting.

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 (MAW), a two-stage framework that enhances long-form legal drafting by decoupling content generation from specific LLM archi-

tectures. 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.

To ensure real-world usability, we developed a Human-in-the-Loop Document Generation System that enables legal professionals to input document types, refine structure, and generate drafts interactively. The system, along with the dataset and code, is made available via an anonymous repository for reproducibility, with plans for public hosting post-acceptance under strict privacy and security guidelines.

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 retrieval-augmented generation, reinforcement learning from human feedback, 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 in this work, several limitations must be addressed in future research. One key constraint is the limited diversity of the training dataset. While `VidhikDastaavej` provides a foundational resource for private legal document generation, it primarily consists of case judgments along with a relatively small set of other legal document types. This imbalance affects the generalizability of `NyayaShilpi`, which struggles with generating underrepresented legal formats. Expanding the dataset to include a more balanced distribution of legal documents such as contracts, agreements, affidavits, and petitions is essential for improving model adaptability.

Another limitation arises from the supervised fine-tuning (SFT) approach on a relatively small dataset. Our findings indicate that `NyayaShilpi` did not exhibit significant improvements after SFT, likely due to the limited number of training examples per legal category. Increasing the volume of annotated legal texts and incorporating additional domain-specific pretraining data could enhance the model's ability to generate diverse and legally accurate documents. Although the Model-Agnostic Wrapper (MAW) significantly improves coherence and logical structuring, it does not entirely eliminate hallucinations in generated legal texts. Some incorrect or irrelevant content may still appear, particularly when the model lacks sufficient contextual grounding. While integrating a retrieval-based mechanism has helped mitigate inconsistencies, additional techniques such as fact verification modules and external legal knowledge sources are required to ensure factual correctness and adherence to legal norms.

A practical limitation of this work is the lack of large-scale real-world deployment and user feedback. While expert evaluations provided insights into factual accuracy and completeness, broader usability testing with practicing lawyers and law firms would offer more comprehensive validation. Assessing the system's adaptability across different legal jurisdictions and case-specific scenarios is crucial before widespread adoption.

Lastly, computational constraints influenced the scope of our experiments. Due to limited resources, fine-tuning was performed on select models, and larger architectures such as LLaMA-3-70B were not explored. Future research should investigate more efficient training techniques, such as parameter-efficient tuning or reinforcement learning, to optimize performance while reducing computational overhead.

Addressing these limitations will be essential for enhancing AI-driven legal document generation, ensuring greater accuracy, reliability, and usability in real-world legal applications.

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, including `NyayaShilpi`, 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. Future work will explore bias-mitigation strategies to enhance fairness in AI-generated legal drafting.

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.

We recognize the accountability challenges in AI-generated legal content. While our tool enhances efficiency, legal responsibility remains with human users, who must review and validate AI-generated drafts before application. To further enhance accountability, future iterations of our system will incorporate traceability features, enabling users to track AI-generated suggestions and modifications.

By addressing these ethical concerns, this work ensures that AI-driven legal tools enhance productivity while upholding privacy, fairness, and professional integrity. The public release of the tool will adhere to copyright, privacy, and security safeguards, ensuring responsible and ethical deployment for legal professionals and researchers.

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.

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.

P Vimala Imogen[1], J Sreenidhi, and V Nivedha. 2024. Ai-powered legal documentation assistant. *Journal of Artificial Intelligence*, 6(2):210–226.

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.

Chun-Hsien Lin and Pu-Jen Cheng. 2024. Legal documents drafting with fine-tuned pre-trained large language model. *arXiv preprint arXiv:2406.04202*.

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.

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.

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*.

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.

Weicong Qin, Zelin Cao, Weijie Yu, Zihua Si, Sirui Chen, and Jun Xu. 2024. Explicitly integrating judgment prediction with legal document retrieval: A law-guided generative approach. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2210–2220, New York, NY, USA. Association for Computing Machinery.

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*.

Jaromír Šavelka and Kevin D Ashley. 2018. Segmenting us court decisions into functional and issue specific parts. In *Legal Knowledge and Information Systems*, pages 111–120. IOS Press.

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.

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.

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.

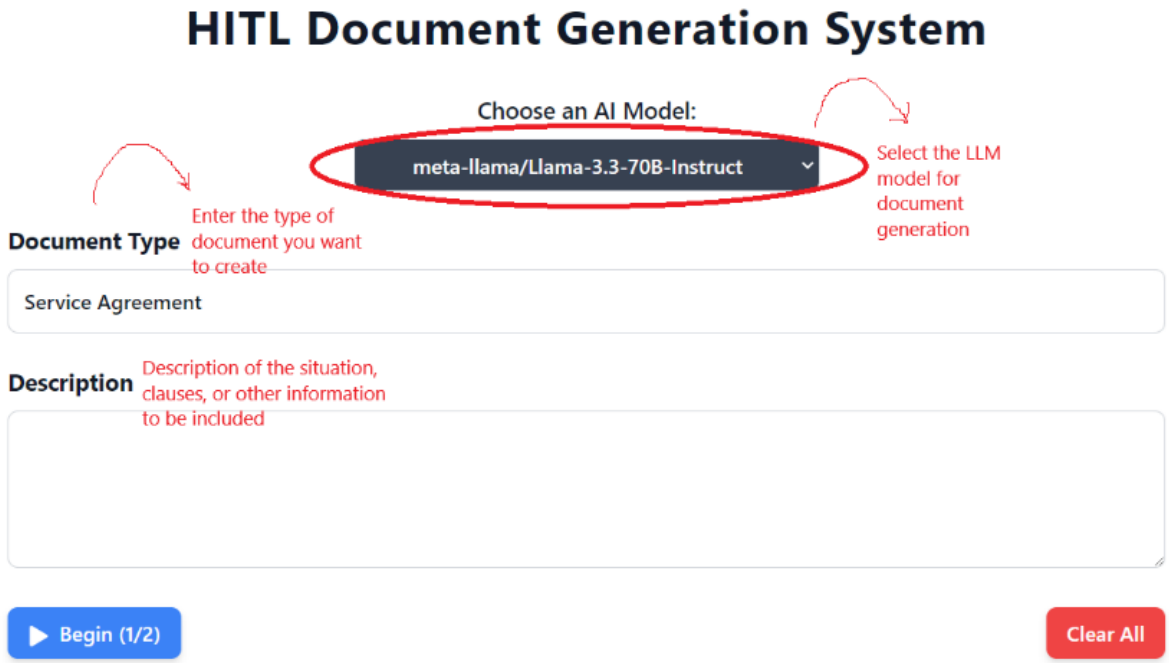

Figure 2: Document Information Entry Interface

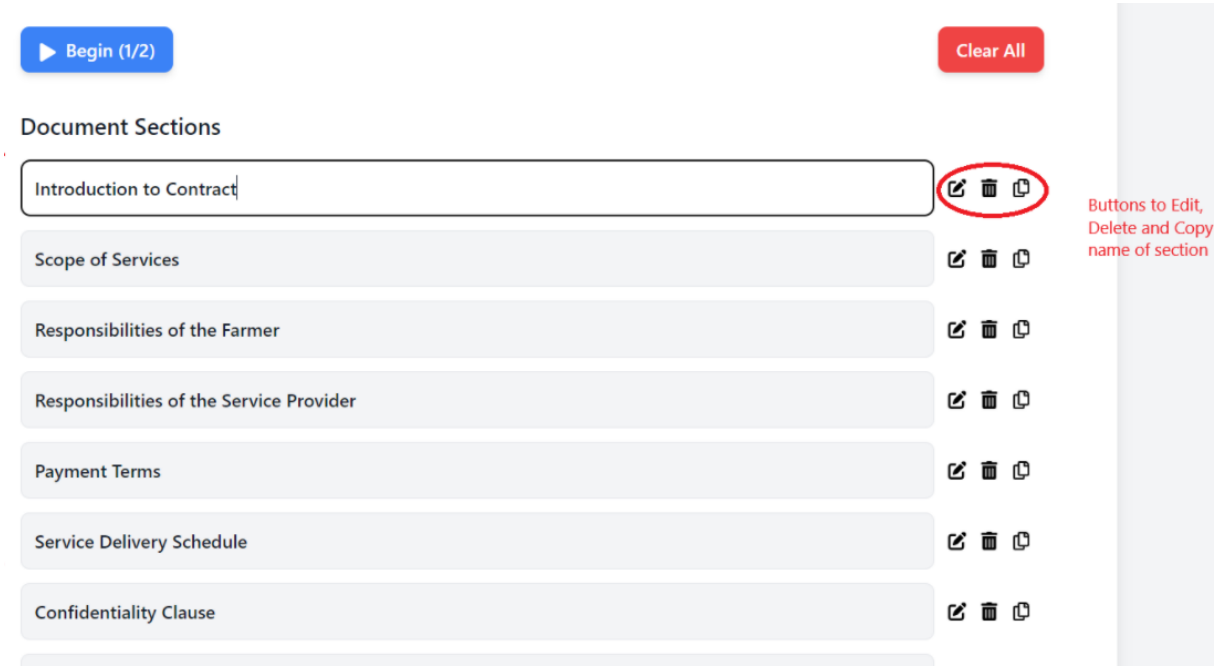

Figure 3: Editing Generated Document Sections

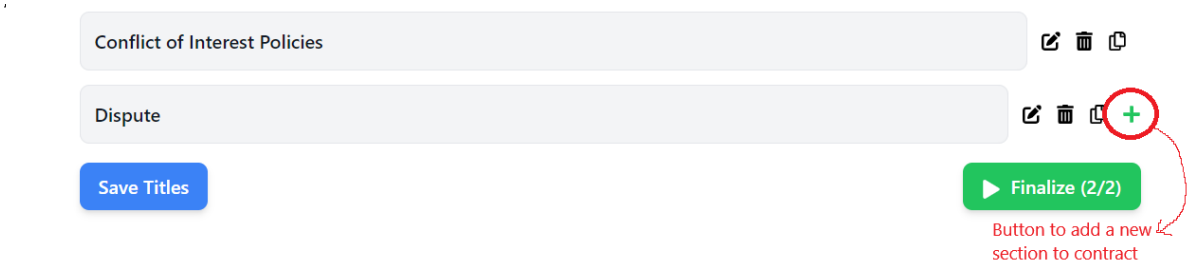

Figure 4: Adding Document Sections

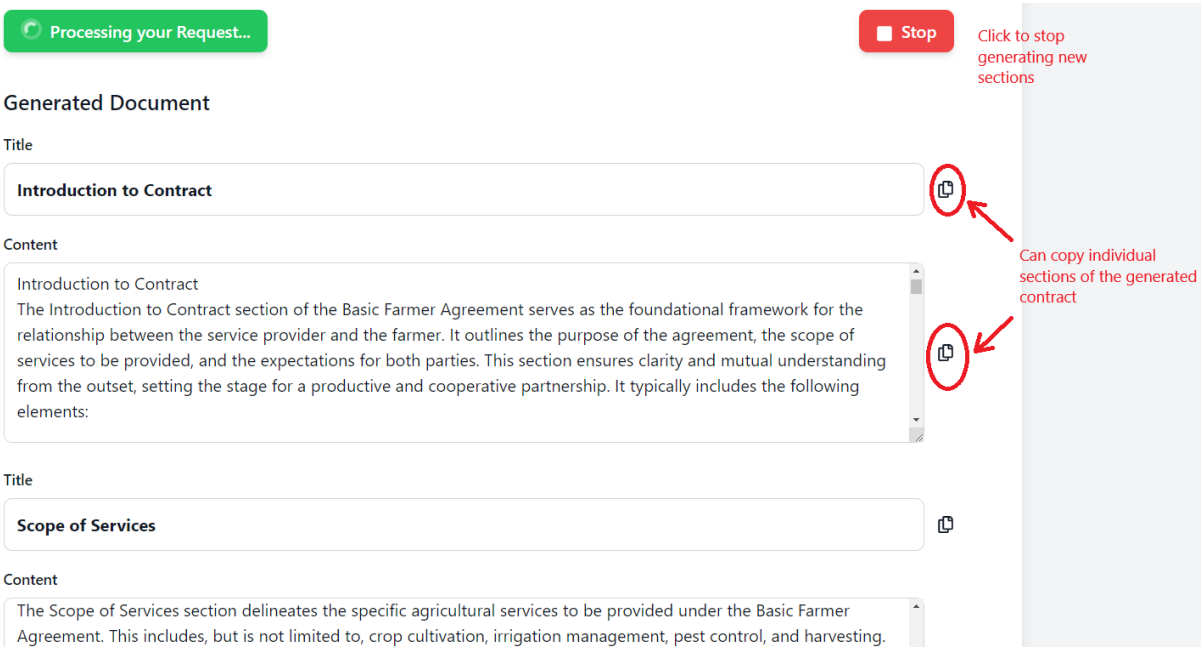

Figure 5: Generating Section Content

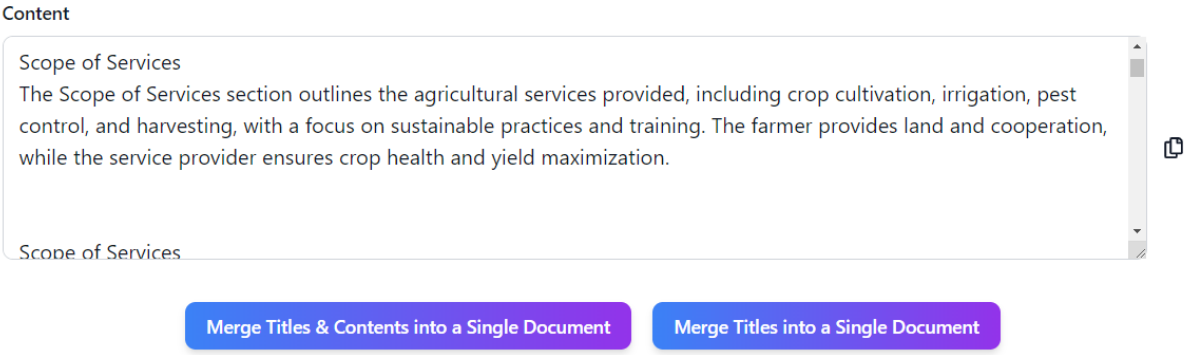

Figure 6: Exporting the Document

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 3: Categories and Corresponding Prompts for Legal Document Generation

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 4: 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 $\rightarrow$ Highly inaccurate, major omissions or distortions, poorly structured.
4-6 $\rightarrow$ Somewhat accurate but incomplete, missing key legal reasoning or context.
7-9 $\rightarrow$ Mostly accurate, well-structured, with minor omissions or inconsistencies.
10 $\rightarrow$ 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 5: 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 6: This table illustrates a sample document after it has been anonymized.