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

Anonymous ACL submission

Abstract

Automating legal document drafting can significantly enhance efficiency, reduce manual effort, and streamline legal workflows. While prior research has explored tasks such as judgment prediction and case summarization, the structured generation of private legal documents in the Indian legal domain remains largely unaddressed. To bridge this gap, we introduce `VidhikDastaavej`, a novel, anonymized dataset of private legal documents, and develop `NyayaShilp`, a fine-tuned legal document generation model specifically adapted to Indian legal texts. We propose a Model-Agnostic Wrapper (MAW), a two-step framework that first generates structured section titles and then iteratively produces content while leveraging retrieval-based mechanisms to ensure coherence and factual accuracy. We benchmark multiple open-source large language models (LLMs), including instruction-tuned and domain-adapted versions, alongside proprietary models for comparison. Our findings indicate that while direct fine-tuning on small datasets does not always yield improvements, our structured wrapper significantly enhances coherence, factual adherence, and overall document quality while mitigating hallucinations. To ensure real-world applicability, we develop a Human-in-the-Loop (HITL) Document Generation System, an interactive user interface that enables users to specify document types, refine section details, and generate structured legal drafts. This tool allows legal professionals and researchers to generate, validate, and refine AI-generated legal documents efficiently. Extensive evaluations, including expert assessments, confirm that our framework achieves high reliability in structured legal drafting. This research establishes a scalable and adaptable foundation for AI-assisted legal drafting in India, offering an effective approach to structured legal document generation.

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.

We propose a Model-Agnostic Wrapper (MAW) framework that structures legal document generation in two phases: section title generation, followed by section-wise content generation. This structured approach enhances coherence, consistency, and factual accuracy, addressing key challenges in long-form text generation. Since `VidhikDastaavej` is relatively small due to the private nature of legal documents, our primary goal is to evaluate the effectiveness of this structured generation approach rather than relying solely on data volume.

To further improve legal text generation, we develop `NyayaShilp`, a domain-adapted model fine-tuned on Indian legal texts. The name `NyayaShilp` is derived from "Nyaya" (justice) and "Shilp" (craftsmanship), emphasizing its ability to craft structured and legally sound documents. The model undergoes two-stage training: (1) Continued pretraining on publicly available Indian legal corpora, such as case laws, to inject domain-specific knowledge, and (2) Supervised fine-tuning on `VidhikDastaavej` to specialize in private legal document generation. We benchmark `NyayaShilp` against both open-source and proprietary LLMs.

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 (Irrelevant) 1–10 (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]. Reviewers can install and run the system locally for validation. 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. `NyayaShilp` Model: A domain-adapted LLM fine-tuned on Indian legal corpora and private legal documents.
3. Model-Agnostic Wrapper: A structured framework ensuring coherence, consistency, and factual accuracy in generated legal drafts.
4. Expert-Based Evaluation Metrics: Introduction of structured legal evaluation focusing on factual accuracy and completeness.
5. 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.,

[1] https://anonymous.4open.science/r/DocGen-887B/

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 knowledge 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	469	20
Number of unique categories	17	7
Avg # of words per document	930.03	818.80
Max # of words per document	3863	1924

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, 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 legal documents, including petitions, legal letters, reply notices, affidavits, lease deeds, memorandums of arguments, compromise petitions, written statements, and bail applications. 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 489 documents, with 469 used for training and 20 for testing. The dataset covers 17 legal document categories in the training set and 7 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 Appendix in Table 3 for the training set and Table 4 for the test set. This diversity is critical for training models that generalize across different legal document formats, improving their usability in real-world legal drafting.

4.2 Document Annotation and Classification

Since the dataset initially lacked labeled document types, we employed an automated classification approach using Mixtral-8-7B-Instruct (Jiang et al., 2024), a model with a 32,000 token context length, to infer document categories. To ensure classification accuracy, a legal expert cross-validated the model's predictions. Test documents were manually annotated by legal experts, ensuring that the evaluation set aligns with actual legal drafting practices.

For training purposes, we generated structured document descriptions as instruction data using Meta-LLaMA-3-70B-Instruct to facilitate model training. In contrast, test document descriptions were manually created by legal experts to ensure that the evaluation dataset remains a robust benchmark for legal document generation.

4.3 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 8.

4.4 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 Methodology

This section describes the key contributions of this work: the development of `NyayaShilp`, a domain-adapted model for legal document generation, and the introduction of a Model-Agnostic Wrapper (MAW) to enhance structured legal drafting.

5.1 Model Training: `NyayaShilp`

5.1.1 Injecting Legal Knowledge

To address the deficiency of Indian legal knowledge in the base LLaMA-2-7B-Chat model, we employed Continued Pretraining (CPT) using a large corpus of Indian legal texts. The training data consisted of 38,321 Supreme Court of India (SCI) cases and 1,00,000 randomly selected High Court cases, ensuring coverage of diverse legal contexts. This step embedded domain-specific knowledge, allowing the model to comprehend legal terminology, argumentation patterns, and document structures. Additionally, the validation set of 12,239 documents from SCI and High Courts was used to monitor adaptation during pretraining., ensuring that `NyayaShilp` effectively captured legal reasoning relevant to Indian jurisprudence.

5.1.2 Task-Specific Fine-Tuning

Following the CPT phase, we conducted Supervised Fine-Tuning (SFT) to specialize the model for private legal document generation tasks. The fine-tuning process utilized a labeled dataset of private legal documents, paired with prompts or situational inputs annotated by legal professionals. These annotations provided detailed mappings of instructions to their corresponding legal drafts, offering high-quality supervision for fine-tuning. For efficient fine-tuning, we employed Low-Rank Adaptation (LoRA) (Hu et al., 2021) to enhance efficiency without requiring full model updates. LoRA provided parameter-efficient adaptation, reducing memory overhead while maintaining task-specific performance.

5.1.3 Final Model: `NyayaShilp`

The final model, `NyayaShilp`, integrates CPT for domain-specific legal knowledge and SFT for structured legal drafting. While designed to enhance the generation of Indian legal documents, our evaluations reveal that direct fine-tuning on a small

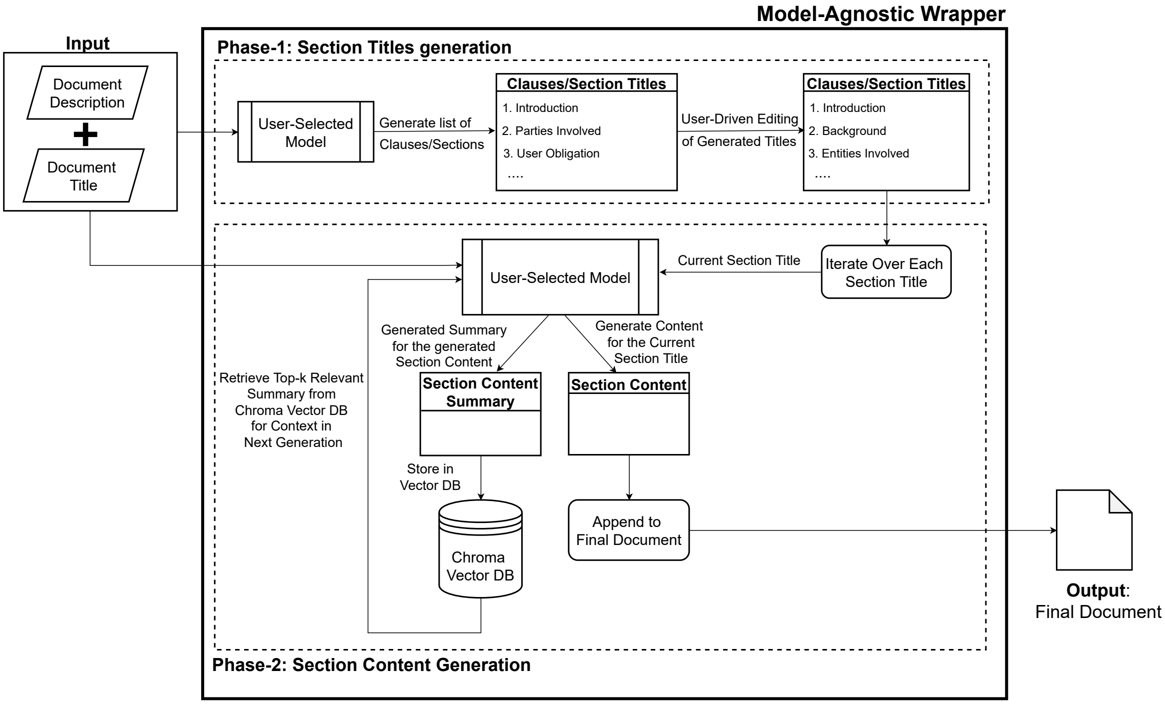


Figure 1: Wrapper flow diagram

dataset does not always yield significant improvements.

6 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.

6.1 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.

7 Experimental Setup

To benchmark the performance of NyayaShilp 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 A.

7.1 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 corpus of Indian legal cases. Further fine-tuning on private legal documents led to LLaMA-2-7B-Chat CPT+SFT (NyayaShilp), 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 5.

7.2 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.

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 7.
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 finetuning (SFT), and the effectiveness of the model-agnostic wrapper.

9.1 Comparative Model Performance

The evaluation results indicate significant variations in performance across different models. Open-source models such as LLaMA-2-7B and LLaMA-3-8B exhibit better performance in lexical and semantic-based evaluations, demonstrating their ability to generate text with high similarity to reference documents. However, GPT-4o, a closed-source model used only for inference, achieves the highest overall scores across multiple metrics, particularly in expert evaluations.

Our instruction-tuned model, NyayaShilp, was expected to enhance factual accuracy and coherence in legal drafting. However, its performance, even after CPT and SFT, did not yield the expected improvements. A potential reason for this could be that the CPT phase primarily included judgment cases and did not incorporate a broader variety of private legal documents, limiting the model's understanding of diverse legal drafting styles. Additionally, across all models, we observed a performance drop after SFT, suggesting that the finetuning dataset might have been insufficiently diverse, particularly in underrepresented categories.

9.2 Impact of Instruction Tuning and Dataset Limitations

Instruction tuning plays a crucial role in adapting general-purpose models for legal document generation. However, our results show that SFT led to performance degradation across models instead of improvement. This indicates that for SFT to be effective, a significantly larger dataset covering all legal document categories is required. Fine-tuning on a limited number of samples per category might have caused overfitting or insufficient generalization, reducing the models' ability to generate consistent and legally sound documents across varied inputs. Some examples of hallucinations encountered in model outputs are provided in Appendix Table 6, due to space constraints.

9.3 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 struc-

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
Phi-3 mini	0.1808	0.0837	0.1203	0.0237	0.0864	0.5074	0.1052	1.9500	0.0000	0.0000
LLaMA-2-7B	0.4439	0.1728	0.2208	**0.0798**	0.2426	0.6225	0.1510	5.5000	5.6643	5.5633
LLaMA-2-7B CPT	0.1563	0.0745	0.1078	0.0121	0.0946	0.5287	0.1152	2.0000	0.1333	0.0583
LLaMA-2-7B CPT+SFT (NyayaShilp)	0.0370	0.0188	0.0252	0.0158	0.0597	0.4798	0.1113	2.1917	0.0000	0.0000
Wrapper (Over LLaMA-2-7B)	0.4436	0.1556	0.2027	0.0518	0.2583	**0.8066**	0.1278	5.1500	6.5547	6.1133
LLaMA-3-8B	0.3154	0.1275	0.1591	0.0552	0.2191	0.6190	0.1593	5.9334	5.9333	5.8417
LLaMA-3-8B SFT	0.0745	0.0359	0.0500	0.0224	0.0729	0.4878	0.1045	2.4094	0.0000	0.0000
Wrapper (Over LLaMA-3-8B)	0.3703	0.1437	0.1756	0.0486	**0.2977**	0.8048	0.1488	6.3834	**8.0667**	**7.5500**
GPT-4o	**0.4506**	**0.1770**	**0.2346**	0.0759	0.2384	0.6241	**0.1599**	**6.5667**	6.0750	6.0750

Table 2: Evaluation Metrics for Different Models. LLaMA-2-7B refers to LLaMA-2-7B-chat, while LLaMA-3-8B refers to LLaMA-3-8B-Instruct. The best results are in bold.

tured, stepwise document generation approach, the wrapper minimizes hallucinations by ensuring that the generated content remains grounded in the given instructions and previously generated sections.

9.4 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 limited category diversity in 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. Further analysis of expert feedback, detailed in Appendix Section D, provides deeper insights into how different models handle legal drafting.

9.5 IAA for Expert Evaluation

To ensure the reliability of expert assessments in evaluating AI-generated legal documents, we conducted an Inter-Annotator Agreement (IAA) analysis. This helps quantify the consistency of human evaluations across models, providing insights into the robustness of our evaluation framework.

We employed Intraclass Correlation Coefficient (ICC) (Koo and Li, 2016), Krippendorff's Alpha (Krippendorff, 2011), and Pearson Correlation Coefficient (Cohen et al., 2009) to measure agreement among three legal experts. These metrics evaluate consistency in factual accuracy and completeness assessments across different models.

Our findings reveal that SFT models and Phi-3 mini exhibit perfect agreement (IAA = 1.00) across all metrics, but this is due to failure in generating meaningful legal drafts, leading to unanimous zero ratings. Baseline LLMs show lower agreement, highlighting inconsistency in outputs, while Wrapper-based models significantly improve agreement, demonstrating structured and coherent legal drafting. GPT-4o, while a strong benchmark, exhibits moderate agreement scores, reinforcing variability in model outputs.

For a detailed breakdown and additional information, see Appendix C, where we present a comprehensive discussion along with Tables 9 and 10.

10 Conclusion and Future Work

This study presents a structured approach to legal document generation using large language models, introducing VidhikDastaavej, a novel dataset of private legal documents, and NyayaShilp, a domain-adapted model fine-tuned on Indian legal texts. We propose a Model-Agnostic Wrapper (MAW), a two-step framework that enhances coherence, logical structuring, and factual accuracy in generated legal documents. Our findings show that continued pretraining and fine-tuning improve domain adaptation, but their impact depends on dataset diversity. NyayaShilp showed limited improvements post-fine-tuning, highlighting the need for a broader dataset covering various legal document types. The MAW significantly enhances long-form legal drafting, ensuring logical structure while minimizing hallucinations. Expert evaluations confirm that wrapper-assisted models perform comparably to GPT-4o, offering a cost-effective alternative for legal drafting. This research advances AI-driven legal drafting, modernizing legal workflows in India.

Future work will expand VidhikDastaavej to include more diverse legal documents, refine fine-tuning strategies for better category balance, and integrate retrieval-augmented generation and reinforcement learning for improved factual accuracy.

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 NyayaShilp, 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 NyayaShilp 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 NyayaShilp, 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.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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 Experimental Setup and Hyperparameters

All experiments were conducted on a single NVIDIA A100 GPU with 40GB memory using the PyTorch framework in conjunction with Hugging Face Transformers. To optimize GPU memory utilization and computational efficiency, mixed precision (fp16) was employed where applicable. Training was further optimized using DeepSpeed Stage 3, enabling memory-efficient distributed training and parameter offloading. Additionally, Low-Rank Adaptation (LoRA) was utilized to perform parameter-efficient fine-tuning, allowing for effective model adaptation without the need for full-scale retraining. Training and evaluation processes were logged using Weights & Biases for detailed tracking, hyperparameter tuning, and loss monitoring.

For instruction tuning, hyperparameters were carefully selected to balance computational efficiency and model performance. Most models were quantized to 4-bit precision to reduce memory consumption while maintaining inference quality. However, CPT LLaMA-2 was trained in 32-bit precision to better capture domain-specific legal knowledge during the continued pretraining (CPT) phase. The maximum sequence length varied across models, with LLaMA-2 supporting up to 2048 tokens and LLaMA-3 allowing sequences up to 4096 tokens, accommodating longer legal texts.

The optimization process utilized AdamW as the optimizer, configured with a learning rate of 1×10^{-4} and a cosine learning rate scheduler, ensuring smooth decay and training stability. To efficiently manage GPU memory while training on large datasets, gradient accumulation was set to achieve an effective batch size of 4.

The fine-tuning (SFT) process was carried out over three epochs for all models. Depending on model size and sequence length, training durations ranged from 38 to 48 hours. LoRA was configured with rank 16, an alpha value of 64, and a dropout rate of 0.1, optimizing memory usage while allowing efficient adaptation of models for private legal document generation.

To facilitate instruction tuning, a set of specialized prompts and instruction sets was created, ensuring the generated legal documents adhered to professional formatting and legal terminology. These instructions helped models learn the structured nature of legal drafts. While two examples of these prompts are provided in Appendix 5, the full instruction set will be made publicly available following the acceptance of this paper to ensure transparency and reproducibility, further advancing research in private legal document generation.

Document Type	**No. of Documents**
Petition	209
Legal Letter	82
Affidavit	44
Memorandum of Arguments	43
Written Statement	25
Bail Application	24
Reply Notice	12
Memorandum of Appeal	8
Legal Notice	5
Lease Deed	4
Vakalatnama	4
Last Will and Testament	3
Memorandum of Understanding	2
Compromise Petition	1
Warrant to the Bailiff	1
MOOT COURT PROBLEM	1
Statutory Application	1
Total	**469**

Table 3: `VidhikDastaavej` Train Doc Type Distribution.

Document Type	**No. of Documents**
Contract	9
Petition	5
Legal Notice	2
Power of Attorney	1
Will and Testament	1
Legal Policy	1
Memorandum of Understanding	1
Total	**20**

Table 4: `VidhikDastaavej` Test Doc Type Distribution.

B HITL Document Generation System: A User Guide

B.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.

B.2 User Interface Guide

B.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.

B.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.

B.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.

B.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.

B.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.

C Inter-Annotator Agreement (IAA) for Expert Evaluation

To ensure the reliability of expert-based evaluation for AI-generated legal documents, we conducted an Inter-Annotator Agreement (IAA) analysis using standard agreement metrics. This evaluation quantifies the consistency of expert assessments in scoring factual accuracy and completeness across different models.

C.1 IAA Metrics and Methodology

We employed three widely used agreement metrics:

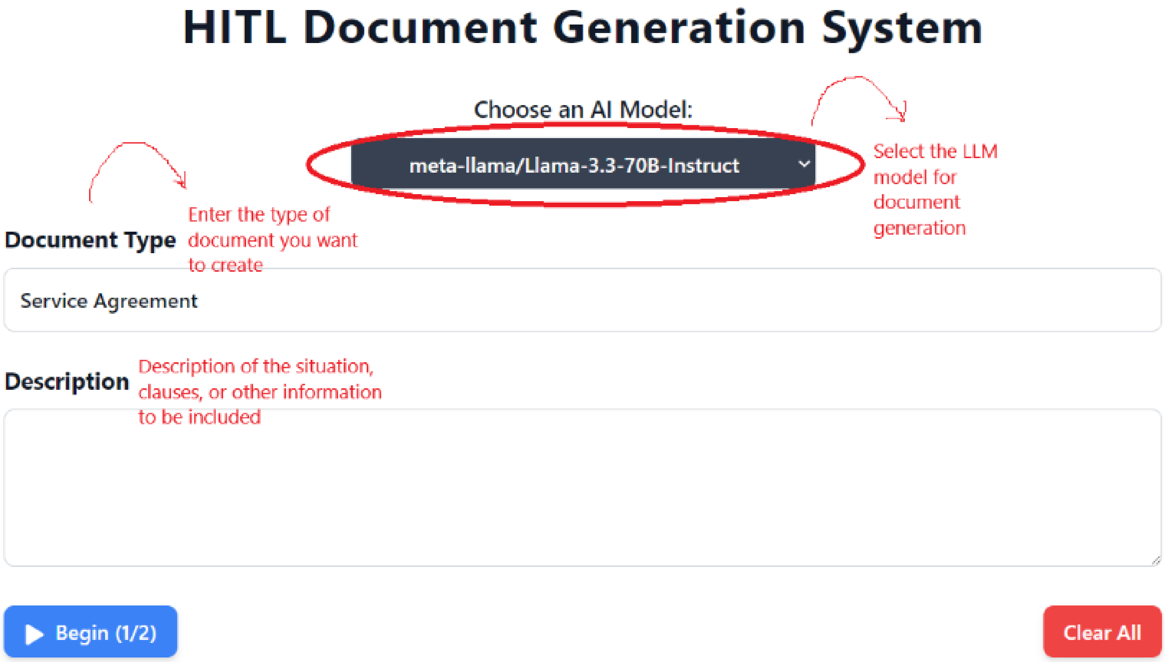
- **Intraclass Correlation Coefficient (ICC)** (Koo and Li, 2016): Measures the absolute agreement among raters for continuous variables, commonly used for reliability assessment in research.
- **Krippendorff’s Alpha** (Krippendorff, 2011): A robust reliability measure applicable to ordinal and interval data, ensuring agreement beyond chance.
- **Pearson Correlation Coefficient** (Cohen et al., 2009): Measures linear correlation between two annotators’ scores, assessing the strength of agreement.

Three legal experts independently rated the generated legal documents for *Factual Accuracy* and *Completeness & Comprehensiveness* using a structured rubric. Each model’s outputs were rated without knowledge of the generating model to prevent bias.

C.2 Findings and Observations

Table 9 presents IAA scores for factual accuracy, while Table 10 provides IAA scores for completeness and comprehensiveness.

High Agreement in SFT and Phi-3 mini Models: The fine-tuned models (`LLaMA-2-7B CPT+SFT` and `LLaMA-3-8B SFT`) and `Phi-3 mini` exhibit perfect agreement (IAA = 1.00) across all metrics.



HITL Document Generation System

Choose an AI Model: meta-llama/Llama-3.3-70B-Instruct

Select the LLM model for document generation

Document Type: Service Agreement

Enter the type of document you want to create

Description

Description of the situation, clauses, or other information to be included

Figure 2: Document Information Entry Interface

However, this is due to their failure in generating meaningful legal drafts, leading to unanimous zero ratings by all experts.

Baseline LLMs Show Moderate Agreement: `LLaMA-2-7B` and `LLaMA-3-8B` exhibit moderate agreement scores, particularly in Pearson correlation. Lower ICC and Krippendorff's Alpha scores indicate inconsistent outputs across different document types, making expert agreement less stable.

Improved Agreement with Wrapper-based Models: Wrapper-based models (`Wrapper (Over LLaMA-2-7B)` and `Wrapper (Over LLaMA-3-8B)`) show significantly higher agreement across all metrics compared to their respective base models. This suggests that the structured, section-wise generation approach improves coherence, making expert ratings more aligned.

GPT-4o Agreement Remains Moderate: While `GPT-4o` demonstrates strong performance in factual accuracy and completeness scores, its IAA scores remain moderate, implying variation in output quality across different document categories.

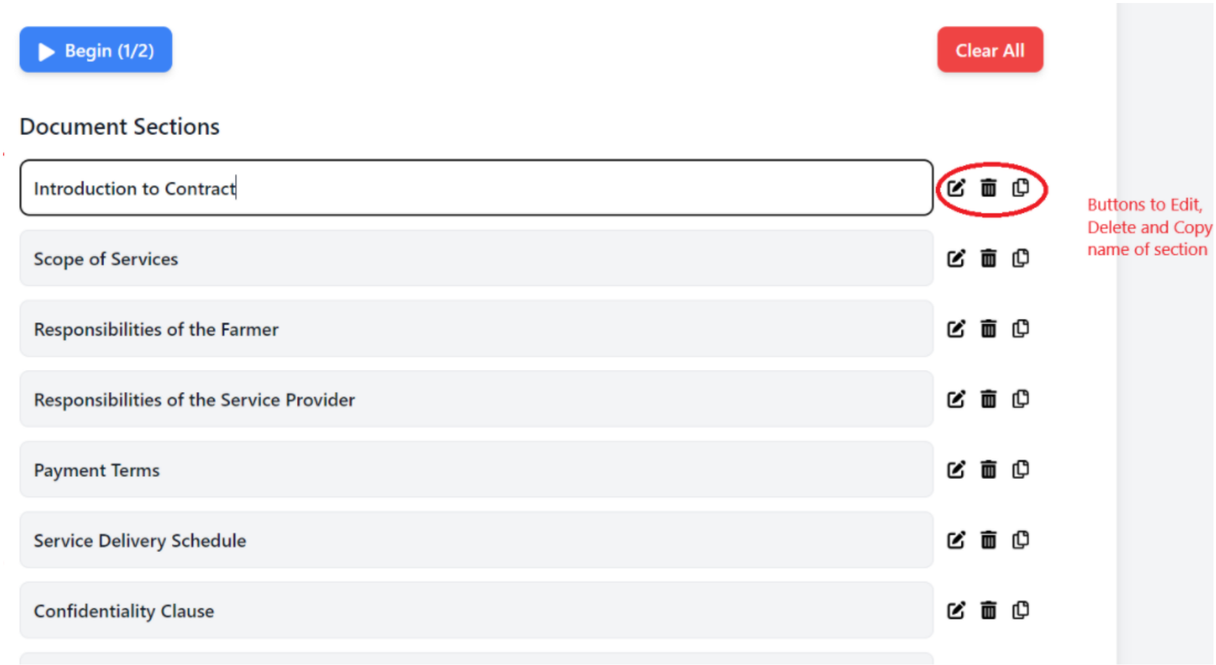
C.3 Conclusion

The results underscore the importance of structured document generation methods, such as the Model-Agnostic Wrapper, in improving consistency in expert evaluations. While base models struggle with consistency, structured approaches enhance coherence and lead to higher expert agreement. The detailed analysis in Tables 9 and 10 supports these findings.

D Insights from Legal Experts

Expert evaluations of the generated legal documents reveal significant variations in model performance across different architectures. Simpler models, such as Phi-3 Mini, fail to produce coherent drafts, frequently hallucinating by inserting random invalid characters or incomplete legal clauses. More sophisticated models, including LLaMA-2-7B Chat and LLaMA-3-8B Instruct, generate well-structured documents but often omit crucial provisions required for legally sound drafts. Experts noted that these models struggle with standard-form contracts, frequently missing key clauses such as jurisdictional provisions, indemnification terms, and policies.

Fine-tuned models exhibited significant limitations. Several instances of hallucination were observed, along with repeated outputs closely mirroring the input prompts rather than generating fully developed legal drafts. This constraint could be introduced by 4-bit quantized fine-tuning, which may have led to reduced model expressiveness, and the limited number of training examples per document type. Some models demonstrated improved fluency but failed to uphold the necessary level of legal rigor, often misrepresenting case law or



Document Sections

Introduction to Contract

Scope of Services

Responsibilities of the Farmer

Responsibilities of the Service Provider

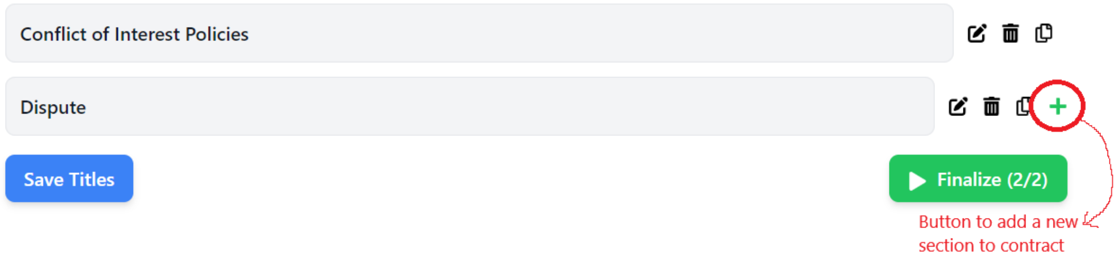
Payment Terms

Service Delivery Schedule

Confidentiality Clause

Buttons to Edit, Delete and Copy name of section

Figure 3: Editing Generated Document Sections


Conflict of Interest Policies

Dispute

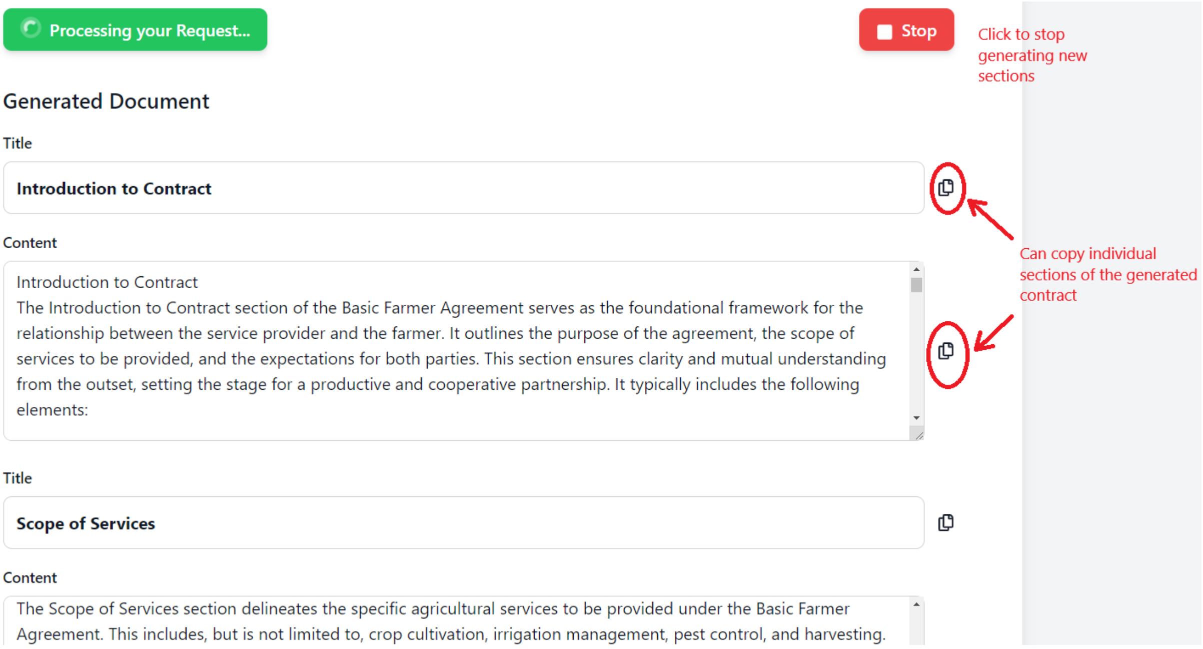
Button to add a new section to contract

Figure 4: Adding Document Sections

contractual obligations.

Evaluators found that GPT-4o generally produced more fluent and formal drafts, yet it exhibited inconsistencies beyond the initial recitals, sometimes failing to address core legal elements in long-form contracts. ChatGPT-generated documents contained simple, readable language but lacked the structured precision expected in professional legal drafting.

The wrapper-based approach improved consistency and ensured that all required sections were present in the generated drafts. However, some experts flagged that the wrapper occasionally introduced extraneous details, adding unnecessary clauses that were not explicitly requested in the input prompt. While this strategy reduced hallucinations and enforced factual accuracy, some reviewers expected more creative legal reasoning rather than rigid compliance with the given instructions.



Click to stop generating new sections

Generated Document

Title

Introduction to Contract

Content

Introduction to Contract
The Introduction to Contract section of the Basic Farmer Agreement serves as the foundational framework for the relationship between the service provider and the farmer. It outlines the purpose of the agreement, the scope of services to be provided, and the expectations for both parties. This section ensures clarity and mutual understanding from the outset, setting the stage for a productive and cooperative partnership. It typically includes the following elements:

Can copy individual sections of the generated contract

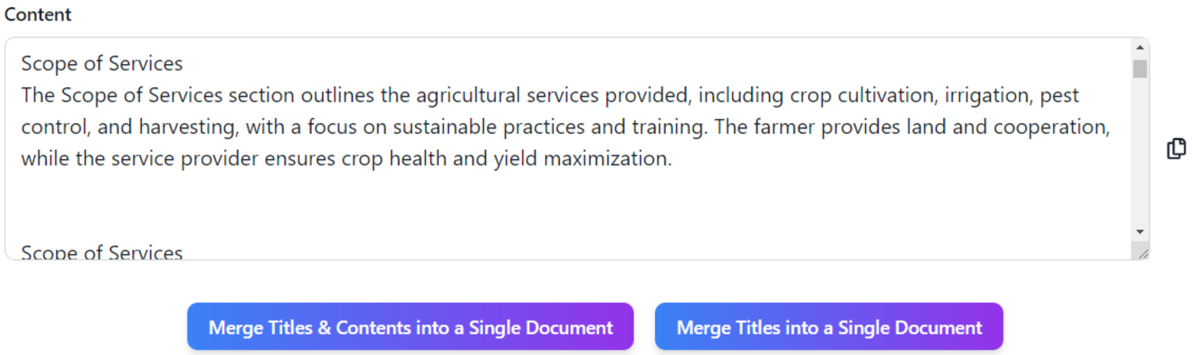
Title

Scope of Services

Content

The Scope of Services section delineates the specific agricultural services to be provided under the Basic Farmer Agreement. This includes, but is not limited to, crop cultivation, irrigation management, pest control, and harvesting.

Figure 5: Generating Section Content


Merge Titles & Contents into a Single Document
Merge Titles into a Single Document

Content

Scope of Services
The Scope of Services section outlines the agricultural services provided, including crop cultivation, irrigation, pest control, and harvesting, with a focus on sustainable practices and training. The farmer provides land and cooperation, while the service provider ensures crop health and yield maximization.

Scope of Services

Figure 6: Exporting the Document

Category	Prompt
Petition	Generate a legal document in the format of a counter filed by a respondent in a civil court case. The document should be titled 'C.M.A. 39/2010' and should be addressed to the 'COURT OF THE CIVIL JUDGE SENIOR DIVISION [PERSON]. The petitioner's name should be '[PERSON]' and the respondent's name should be 'Mr. [PERSON]'. The counter should state that the petition filed by the petitioner is based on false pleas and lacks substantial proof, and that all the averments made in paragraphs [CARDINAL] and [CARDINAL] of the petition are false and fabricated. The document should also mention that the petitioner failed to adduce evidence to support their claims despite being given sufficient time, and that the present petition was filed belatedly, causing delay and wasting the court's valuable time. The counter should also reference an order dated [DATE] in [GPE] 219/2006, which dismissed the suit after giving the petitioner ample opportunities to present their case. Finally, the document should conclude with a prayer to the court to dismiss the [ORG] in the interest of justice, and should be signed by the advocate for the respondent. The tone of the document should be formal and professional, with a slightly argumentative tone in response to the petitioner's claims. The language should be precise and concise, with proper use of legal terminology and formatting.
Legal Letter	Generate a formal letter addressed to the [ORG], Energy [GPE], [GPE], in response to a notice issued on [DATE], vide [PERSON] [DATE], dated [DATE], under [LAW] of the Indian Electricity Act [DATE]. The letter should be written from the perspective of [GPE], son of [PERSON], an engineer residing at [PERSON] [ORDINAL] line, [GPE] [PERSON], [GPE]. The letter should include a detailed explanation in response to the allegations made against [GPE] by [ORG], Bijipur-1, [PERSON]. Specifically, the letter should deny the allegations of tampering with the meter on [DATE], and claim that the allegations are baseless and made with malice. The letter should also mention that [GPE] was not present during the visit by the [ORG] survey Staff and was not informed when they took the meter from the affixed place. Additionally, the letter should state that [GPE] has preferred an application before the Permanent Lok [PERSON] at [GPE], which has been registered as [GPE] case No. 91/2012, and that the [ORG] Chairman and Members have directed the [ORG] to restore the power supply immediately. The letter should conclude by requesting that no action be taken against [GPE] based on the report presented by [ORG], Bijipur-1, and that a clean enquiry be conducted to investigate the illegal entry of the workers sent by S.D.O. Bijipur-1. The letter should be signed by [FAC] and include his full name and address. The tone of the letter should be formal and professional, with a hint of assertiveness and defensiveness.

Table 5: 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`
		`Tortildenesel~ildeildenild`
		`enajseselUNEomor`
		`~ilde~`
		`Tort~ajsague~eselilden~`
		`~agogueUNE~ibbon~ attentesel~`
		`ULKLEV tortomoreltas~ilde~`
		`agateildeinasULKagenUNE~`
		`ildeildeilde~`
		`inflictilenamesildeildeagen~`
		`LEVLEVULK~ildeinasLEVilNE`
		`TortildeLEVildeagogue~`
		`ildeagateilden~ilde~ ~`
		`tortteNELEVinishedULK~ildeinas`
		`Tort`
		`attentLEV~ildenLEVLEV~teg`
		`TortLEV attentLEV~ ~agate`
		`attent tort attent~ildeULK`
		`~ULKULKagogueagateldenULK`
		`attentildeLEVULK~ULK`
		`tortUNEesonildeULK`

Table 6: Example of hallucinations in AI-generated (LLaMA-3-8B-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.

Models	**Intraclass Correlation**	**Krippendorff's**	**Pearson Correlation**		
	Coefficient (ICC)	**Alpha**	**Expert1 vs Expert2**	**Expert1 vs Expert3**	**Expert2 vs Expert3**
Phi-3 mini	1.0000	1.0000	1.0000	1.0000	1.0000
LLaMA-2-7B	0.4833	0.2818	0.7966	0.7339	0.9373
LLaMA-2-7B CPT	0.4865	-0.0064	0.3267	0.1961	0.9337
LLaMA-2-7B CPT + SFT	1.0000	1.0000	1.0000	1.0000	1.0000
Wrapper (Over LLaMA-2-7B)	0.6689	0.1240	0.8957	0.9356	0.9248
LLaMA-3-8B	0.1521	0.4954	0.1408	0.0420	0.9354
LLaMA-3-8B SFT	1.0000	1.0000	1.0000	1.0000	1.0000
Wrapper (Over LLaMA-3-8B)	0.7652	0.0990	0.9422	0.9216	0.9388
GPT-4o	0.1343	0.4489	0.0150	0.0826	0.9320

Table 9: Inter-Annotator Agreement (IAA) Metrics for Factual Accuracy, evaluating consistency among expert reviewers across different models.

Models	**Intraclass Correlation**	**Krippendorff's**	**Pearson Correlation**		
	Coefficient (ICC)	**Alpha**	**Expert1 vs Expert2**	**Expert1 vs Expert3**	**Expert2 vs Expert3**
Phi-3 mini	1.0000	1.0000	1.0000	1.0000	1.0000
LLaMA-2-7B	0.5140	0.1901	0.7234	0.6967	0.9106
LLaMA-2-7B CPT	0.2785	-0.0153	-0.0526	-0.0765	0.6882
LLaMA-2-7B CPT + SFT	1.0000	1.0000	1.0000	1.0000	1.0000
Wrapper (Over LLaMA-2-7B)	0.8356	0.0195	0.9561	0.8845	0.9432
LLaMA-3-8B	0.1837	0.3299	0.0888	-0.0034	0.9145
LLaMA-3-8B SFT	1.0000	1.0000	1.0000	1.0000	1.0000
Wrapper (Over LLaMA-3-8B)	0.8299	0.0178	0.9453	0.8719	0.9370
GPT-4o	0.1382	0.3219	0.1004	-0.0579	0.9047

Table 10: Inter-Annotator Agreement (IAA) Metrics for Completeness & Comprehensiveness, evaluating consistency among expert reviewers across different models.