InLegalLaySum: A Novel Benchmark for Lay Summarization of Indian Legal Documents

Trishita Saha[*1] **Debtanu Datta**[*1]
Shubham Tiwari[1] **Mohit Kaushal**[2] **Priyank Rathore**[2]
Mohammad Khalid Pandit[2] **Pawan Goyal**[1] **Saptarshi Ghosh**[1]

[1]IIT Kharagpur, India [2]NIT Hamirpur, India
Correspondence: trishitasaha.kgp@gmail.com

Abstract

Legal documents are often lengthy and contain domain-specific terminology and complex legal reasoning, making them difficult for the general public to understand. This challenge is further amplified in the Indian context, where a large portion of the population lacks proficiency in complex English. Although automatic legal summarization has been studied, most existing approaches generate summaries intended for legal professionals and thus remain difficult for laypeople. To address this gap, we introduce **InLegalLaySum**, the first benchmark for lay summarization of Indian legal documents. InLegalLaySum comprises 9,570 judgements from the Supreme Court and several High Courts of India paired with expert-written summaries in English. We benchmark a wide range of generative models (ranging from 160M to 7B parameters) and perform comprehensive experiments with multiple training strategies, including supervised fine-tuning (SFT) and direct preference optimization (DPO). We observed that the hybrid SFT+DPO training achieved superior performance across readability metrics, which is further validated by domain experts.

1 Introduction

Legal documents, particularly court judgements, are often lengthy and drafted in complex legal language that includes domain-specific terminology, detailed procedural descriptions, and intricate legal reasoning. Such characteristics make these documents difficult for the general public to understand. While automatic legal summarization (L-Summ) of court judgements has long been explored (Shukla et al., 2022; Malik et al., 2024), existing approaches primarily focus on generating summaries suitable for legal professionals, which are not useful to laypeople who lack formal legal training. Therefore, individuals seeking to understand the outcomes and implications of legal decisions often struggle to interpret the content. Consequently, *there is a pressing need for models that can not only summarize long, complex legal judgements but also present the essential information in a clear, comprehensible manner to laypeople* (people without professional legal training)[1]. To address this gap, we introduce **InLegalLaySum** (**In**dian **Legal Lay Sum**marization), a dataset comprising 9,570 judgements from the Supreme Court and several High Courts of India, paired with expert-written summaries in English. We also provide synthetically generated simplified summaries for training models for the lay summarization task To the best of our knowledge, this is the *first* dataset designed to address the lay summarization task in the Indian legal domain. An example illustrating this legal lay summarization task is shown in Figure 1.

Using InLegalLaySum, we benchmark a variety of both encoder-decoder and decoder-only generative models ranging from 160M to 7B parameters. In addition to open-source models, we also evaluate the state-of-the-art (SOTA) proprietary large language model (LLM) *GPT-5*. We employ multiple training strategies, including supervised fine-tuning (SFT) and direct preference optimization (DPO). Furthermore, our evaluation framework includes both automatic metrics and analysis conducted by domain experts and laypeople. To summarize, the contributions are as follows:

- We introduce **InLegalLaySum**, the first benchmark dataset for lay summarization in the Indian legal domain, comprising 9,570 Supreme Court and High Court judgements paired with expert-written summaries in English. We also provide synthetically generated simplified summaries for training models for the lay summarization task.
- We evaluate a wide range of encoder-decoder and decoder-only generative models, along with

[*]Equal contribution by the first two authors.

[1]In this study, we refer to the term 'Laypeople' to indicate a person with up to a high school (12th Grade) education.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3099–3123
November 6-10, 2026 ©2026 Association for Computational Linguistics

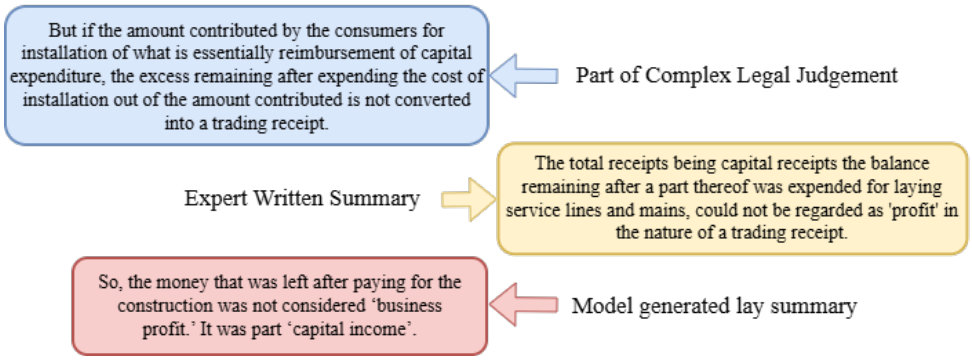


Figure 1: Example illustrating the difference between a complex legal judgement, an expert-written summary, and a model-generated lay summary.

GPT-5, for the legal lay summarization task.

- We investigate multiple training paradigms, including SFT and DPO, to improve the generation of legal lay summaries.
- We conduct extensive evaluation using automatic metrics, as well as evaluations by domain experts and laypeople, focusing on readability, faithfulness, and usefulness for laypeople.

The dataset and codes of this paper are available at https://github.com/Law-AI/InLegalLaySum.

2 Related Works

In this section, we discuss existing works on lay summarization, text simplification, and legal summarization in the Indian jurisdiction.

Text Simplification. The aim of text simplification is to convert complex text into a simpler text while preserving the original meaning. The current research investigates two aspects of text simplification: modeling approaches and their evaluation methods. *BLESS* (Kew et al., 2023) evaluated LLMs by testing their ability to simplify sentences and found that LLMs generate fluent text, yet they produce semantic changes and unauthentic alterations. Traditional automatic readability metrics are commonly used for measuring simplification, especially *Gunning Fog Index (GFI)*[2] and *Flesch-Kincaid Grade Level (FKGL)*[3]. Both measure text readability based on word complexity and sentence structure, with *lower scores indicating higher readability*. Existing simplification works commonly adopt these metrics to measure readability (Martin et al., 2020). Other metrics for text readability include *Dale–Chall (DC)*[4] and *Coleman–Liau index (CLI)*[5] are also used following prior readability-aware summarization work (Ribeiro et al., 2023).

Summarization in Indian Legal Domain. Existing research on summarization in the Indian legal domain has focused on developing several resources. (Shukla et al., 2022) released the *IN-Abs* dataset, which contains Indian Supreme Court judgments paired with abstractive summaries in English. Datta et al. (2023) introduced *MILDSum*, a multilingual L-Summ dataset that contains Indian Court judgements in English, along with their summaries in both English and Hindi. More recently, (Malik et al., 2024) released *CivilSum*, a large-scale summarization dataset focused on Indian civil court judgements and their summaries in English. However, all these datasets contain summaries suitable for Law practitioners, *not* for laypeople.

Lay Summarization in Non-legal Domain. The aim of lay summarization is to reduce both the linguistic complexity and length of text to make it easily comprehensible for non-expert readers. Existing research on lay summarization focused on the scientific and biomedical domains. The *BioLaySumm Shared Task* (Goldsack et al., 2023) introduced a dataset for generating lay summaries of biomedical articles, producing concise explanations that are easy for the general public to understand. Further work extended by investigating improved modeling strategies and evaluation methods for biomedical lay summarization (Ming et al., 2025).

Despite these developments, research on lay summarization has been limited to the scientific and biomedical domains. Though prior work on

[2] https://en.wikipedia.org/wiki/Gunning_fog_index

[3] https://en.wikipedia.org/wiki/Flesch-Kincaid_readability_tests

[4] https://en.wikipedia.org/wiki/Dale_Chall_readability_formula

[5] https://en.wikipedia.org/wiki/Coleman_Liau_index

Split	#Datapoint	Category	Avg #Tokens	Readability Scores			
				GFI↓	FKGL↓	DC↓	CLI↓
Train	7470	Judgement	5720	12.47	12.33	7.39	10.19
		Expert Summary	1018	14.84	13.89	8.66	10.64
		Simplified Summary	805	9.15	8.44	7.39	9.37
Test_Complex	421	Judgement	6356	13.35	13.24	7.35	10.00
		Expert Summary	1082	10.61	9.66	8.43	9.52
Test_Simpler	1679	Judgement	4645	10.22	9.82	7.35	9.61
		Expert Summary	952	10.43	9.47	8.31	9.69

Table 1: Dataset statistics. Lower readability scores indicate simpler and more understandable text.

lay summarization (Moro et al., 2025) exits but to the best of our knowledge, there is no prior work on lay summarization in the Indian legal domain. This gap is crucial because Indian legal judgements are complex, making them difficult for the general public to understand. In this study, we address this limitation by introducing `InLegalLaySum` and experimenting with several strategies for lay summarization of Indian legal judgements.

3 The `InLegalLaySum` Dataset

In this work, we develop **`InLegalLaySum`** (**In**dian **Legal Lay Sum**marization), a large-scale benchmark comprising 9,570 judgements from the Supreme Court and several High Courts of India, along with their abstractive summaries in English. To our knowledge, `InLegalLaySum` is the first benchmark for legal summarization for laypeople within Indian jurisdiction. Table 1 gives several statistics of the dataset.

Dataset Curation. Constructing a dataset for legal lay summarization is *challenging because publicly available summaries of court judgements are typically written by domain experts*, where the end-users (readers) are primarily law professionals. Further complexities arise from the *time-intensive nature of summarizing lengthy legal judgements* and the *limited availability of domain experts*. To address these challenges, we leverage two existing `L-Summ` datasets from the Indian jurisdiction – ***IN-Abs*** and ***MILDSum*** (described in Section 2). Additionally, we curate samples from *BarAndBench*[6] (a popular legal news website among law professionals), which publishes news articles summarizing Indian court judgements. The derivation of lay summaries from the *BarAndBench* source is explained in detail in Appendix A.1. The `InLegalLaySum` dataset consists of 2 splits: *train* (7470 datapoints) and *test* (2100 datapoints). We describe the strategy for constructing these splits below.

Test set. To construct the *test* split, we select only those datapoints whose expert-written summaries are suitable for laypeople (i.e., individuals with up to high school or 12th grade education), as determined by the *FKGL* and *GFI* readability metrics (both ≤ 12). The choice of metrics used to create the test set is explained in Appendix A.3. Specifically, we include 1,512 datapoints from *IN-Abs*, 442 datapoints from *MILDSum*, and 146 datapoints from *BarAndBench* that satisfy this readability criterion, i.e., for which the expert-written summaries are comprehensible to laypeople. In total, the resulting *test* split comprises 2,100 datapoints.

Quality and splits of the Test set. We specifically included in the test set those datapoints where the expert-written summaries are simpler (having FKGL and GFI score ≤ 12). So, a question naturally arises – does the test set consist only of simpler judgements? To investigate this, we compare the readability metrics of the 2,100 judgements in our test set with those of several standard judgement datasets from the Indian legal domain. Figure 2 compares the distribution of FKGL and GFI metrics of the judgements in our test set (`InLegalLaySum _Test`) and the judgements in several Indian legal datasets: ILC (Trivedi et al., 2024), CivilSum (Malik et al., 2024), and MILDSum (Datta et al., 2023). We see that the distribution of document complexity levels in our test set is quite similar to that of the other datasets. Hence, although we constructed the test set by selecting documents whose expert-written summaries are simple, the judgements in our test set span a wide range of complexity levels, without any bias towards simple judgements.

[6] https://www.barandbench.com/

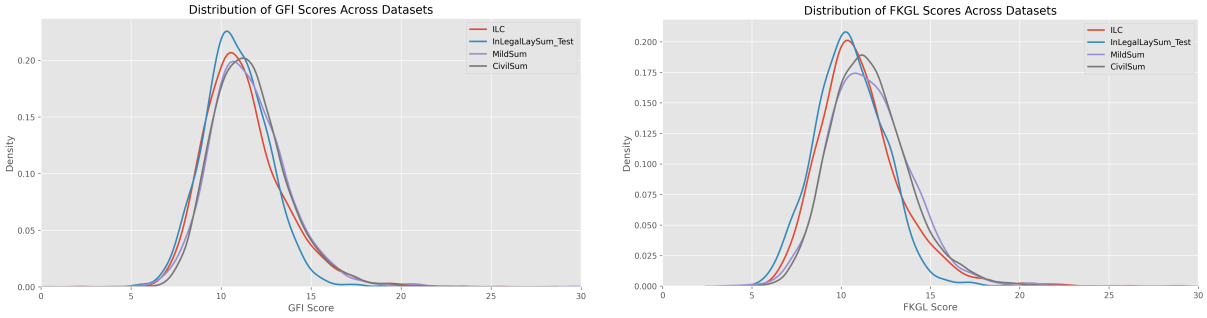


Figure 2: Distribution of GFI and FKGL metrics across `InLegalLaySum` _Test and different Indian legal datasets. `InLegalLaySum` _Test is very similar to existing Indian legal datasets in terms of the legal complexity of judgements.

Further, to benchmark the models over judgements having different difficulty levels, we divide our test set into two subsets:
(i) *Test_Complex*: complex judgements having both GFI and FKGL scores > 12 (421 datapoints), and
(ii) *Test_Simpler*: simpler judgements having both GFI and FKGL scores ≤ 12 (1679 datapoints).
Dataset statistics for both these test splits are provided in Table 1. Detailed statistics of the distribution of lengths of samples in `InLegalLaySum` are described in Appendix A.2.

Train set. The *train* split is constructed from the remaining data (leaving out the 2,100 samples in the test set), where expert-written summaries are more complex according to the readability metrics (either *FKGL* or *GFI* is > 12). Next, to train models for the lay summarization task, we synthetically prepared simplified versions of these training summaries as described in the next section.

4 Lay Summaries for Training Models

As discussed in Section 3, the expert-written summaries in *train* split of `InLegalLaySum` are complex for laypeople according to the readability metrics. Therefore, to train the models for the lay summarization task, we need a simplified version of the summaries in the *train* set that is comprehensible to laypeople. Now, manually drafting a large-scale training set with lay summaries is highly challenging and expensive, especially in high-stakes domains like law. To address this, we consult with domain experts and follow prior work on lay summarization (Moro et al., 2025), and utilize a SOTA LLM to prepare a simplified version of the expert-written summaries in the train set, which are used exclusively for training.

Choice of LLM for synthetic training summaries. To choose the LLM for simplifying training summaries, we conduct an experiment on a subset of `InLegalLaySum`: 700 training samples (~10% of the train set) and 200 test samples (~10% of the test set). First, we generate two sets of simplified summaries of the same expert-written summaries from this train subset with *Gemini-2.0-Flash* (*Gemini*) and *GPT-5.4-mini* (*GPT*), yielding two parallel training sets: *Train-Gemini* and *Train-GPT*. The LLMs are prompted to simplify the expert-written summaries so that the resulting summaries are comprehensible to laypeople. The prompt for generating simplified version of the expert-written training summary is shown in Appendix B (Figure 8).

Next, we train Llama and Gemma using both SFT and SFT+DPO approaches (described in Section 5.2) on each of these two training sets, and apply them on the test subset (containing 200 test samples). The results in Table 2 show similar trends in performance. Although GPT-simplified summaries lead to higher scores on most quality metrics (described in Section 6.1), the Gemini-simplified summaries result in better readability while maintaining comparable overall performance. This balance makes Gemini a suitable choice for generating simplified summaries for the entire *train* set of `InLegalLaySum`.

Example of an expert-written summary and its simplified version by Gemini are given in Figure 9 and Figure 10, respectively, in Appendix A.4. We also make the Gemini-generated simplified summaries (from the train split of `InLegalLaySum`) available as part of our dataset.

Quality of synthetic training summaries. To ensure the quality of *Gemini*-simplified training summaries, we conducted an expert evaluation of 100 random (expert-written, *Gemini*-simplified) summary pairs from the train set. One author, who is an LL.M. (Master of Laws) graduate from a reputed Law school in India, rated each pair on a 5-point

Models	Method	Simplification Model	Readability				Quality				Simplification
			GFI↓	FKGL↓	DC↓	CLI↓	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	SARI
Llama	ZSI	-	9.12	8.54	7.49	9.53	39.41	11.22	17.24	82.42	34.18
	SFT	Gemini	8.69	7.83	**7.74**	**8.13**	43.66	15.85	19.67	83.07	35.67
		GPT	**8.65**	**7.79**	8.17	8.35	**45.33**	**17.62**	**20.20**	**83.33**	**36.06**
	SFT+DPO	Gemini	7.76	6.74	**7.09**	6.89	35.51	8.85	16.58	81.94	33.59
		GPT	**7.50**	**6.09**	7.69	**6.39**	**39.17**	**11.96**	**17.56**	**82.65**	**34.39**
Gemma	ZSI	-	9.41	7.98	8.11	9.34	27.39	6.89	14.12	78.83	31.61
	SFT	Gemini	**8.09**	**7.47**	**7.24**	8.47	41.35	13.37	18.88	82.53	34.58
		GPT	8.83	8.04	8.00	**8.45**	**44.57**	**15.63**	**19.96**	**83.13**	**35.24**
	SFT+DPO	Gemini	**7.33**	**6.18**	**6.90**	**6.95**	36.61	9.29	17.66	82.01	33.07
		GPT	7.79	6.37	7.04	6.99	**40.52**	**11.63**	**18.89**	**82.87**	**33.71**

Table 2: Results of experimental methods where lay summaries are created by both Gemini and GPT.

Metric	Description
Relevance & Coverage (RC)	How well does the summary capture the important information from the judgement? How much of the important content from the judgement is covered in the summary?
Factual Consistency (FC)	Are all facts in the summary consistent with the original judgement?
Legal Terms Explained (LTE)	Explanation of legal maxims and legal terminology.
Grammaticality (Gr)	Is the summary grammatically correct?
Coherence (Co)	Are ideas presented in a logical and connected manner?
Conciseness & Hallucination (CH)	Is the summary free of unnecessary details, verbose language or hallucinated content?

Table 3: Description of metrics used for expert evaluation

Metric	Average Score
Relevance & Coverage (RC)	4.97
Factual Consistency (FC)	4.89
Legal Terms Explained (LTE)	4.52
Grammaticality (Gr)	4.88
Coherence (Co)	4.99
Conciseness & Hallucination (CH)	4.76

Table 4: Average expert evaluation scores over 100 random samples from the *train* split of `InLegalLaySum`.

Likert scale (1: lowest; 5: highest) according to six mutually discussed metrics (Table 3) and provided overall remarks. The average score across all metrics is 4.84 out of 5, indicating the high quality of *Gemini*-simplified training summaries. The detailed results are shown in Table 4. Analysis of training data is provided in Appendix A.4 (Table 11), including qualitative analysis.

5 Experiments

In this section, we describe the models and methods that we considered in this study, along with the details of the experimental setup.

5.1 Models

We evaluate a wide range of encoder-decoder and decoder-only generative models, ranging from 160M to 7B parameters. We also benchmark the SOTA proprietary LLM, *GPT-5* for legal lay summarization. The list of the models we benchmark is given below. The detailed descriptions are in Appendix C.

(1) **Legal-LED-Base-16384** (Legal LED): A 162M parameter encoder-decoder model, specifically designed for long legal text. (2) **Legal Pegasus**: A 560M parameter encoder-decoder model having a context window of 1024, meant for abstractive summarization of legal documents. (3) **Flan-T5-Large** (Flan T5): A 780M parameter encoder-decoder model having a context window of 512. (4) **Gemma-2B-IT** (Gemma): A 2B instruction-tuned decoder-only LLM having a context window of 8K. (5) **Llama-3.2-3B-Instruct** (Llama): A 3B instruction-tuned decoder-only LLM having a context window of 128K. (6) **Phi-4-mini-instruct** (Phi4): A 4B parameter instruction-tuned decoder-only model having a context window of 128K. (7) **Mistral-7B-Instruct-v0.3** (Mistral): A 7B parameter instruction-tuned decoder-only model. (8) **GPT-5**: An instruction-tuned decoder-only proprietary SOTA LLM for advanced text generation.

5.2 Training strategies for models

We investigate diverse strategies for summarization and simplification, which are described below.

Supervised Fine-Tuning (SFT). First, we perform supervised fine-tuning with all the encoder-decoder and decoder-only models described earlier. While fine-tuning, the *Gemini*-generated simplified summaries are considered as the target output of the given judgments.

Direct Preference Optimization (DPO).

DPO (Rafailov et al., 2023) is a training method that optimizes language models using preference data to align model outputs with preferred responses. We conduct DPO experiments with the decoder-only models described earlier. Here, the expert-written summaries (with high FKGL & GFI scores) of the *train* set are considered as the rejected summaries, while the *Gemini*-generated simplified summaries are considered as the preferred summaries. The preference signal is synthetic rather than human-annotated, but this is a deliberate and necessary design choice. As described in Section 4, manually drafting large-scale legal lay summary pairs by domain experts is expensive.

Hybrid Strategy (SFT + DPO). Following Mullick et al. (2025), we adopt a hybrid strategy that combines SFT with preference-based alignment in a two-stage training pipeline. SFT is first performed on half of the training samples to ensure stable legal domain adaptation and faithful content preservation. The SFT-initialized model is then further optimized using DPO on the remaining training samples, which encourages a higher likelihood of preferred summaries (simplified summaries) over non-preferred summaries (complex summaries). This two-stage training pipeline first establishes a faithful representation through SFT and subsequently refines outputs to improve readability via preference alignment (DPO).

Zero Shot Inference (ZSI_Direct). As a baseline, we perform zero-shot off-the-shelf inference over the *test* split, to generate legal lay summaries without any task-specific training. The full prompt used here is shown in Appendix B (Figure 11).

Summarize-Then-Simplify Pipeline (ZSI_Pipeline). Here we first perform summarization of a document and and then perform simplification on the summary. Both steps are performed via zero-shot off-the-shelf inference. The prompts used for summarization and simplification in the ZSI_Pipeline are shown in Appendix B (Figure 12 and Figure 13).

5.3 Experimental Setup

Chunking of long documents. Some models, like Flan T5 and Legal Pegasus, have input capacities of 512 and 1024 tokens, respectively. Hence, a case judgement may not fit fully into the model, and we adopt a chunking strategy for these models. More details are discussed in Appendix D.

Hyperparameter & Infrastructure. The experiments were carried out on a system having an NVIDIA L40 GPU (48 GB vRAM). We use the default hyperparameters to benchmark their performance, and for SFT we adopt parameter-efficient fine-tuning (PEFT) with LoRA hyperparameters ($r = 8$, $\alpha = 32$, dropout=0.05) and adapters injected into the attention and feed-forward projection layers. We use the *AdamW* optimizer with 10 epochs and a learning rate of $2e-5$. We use greedy decoding, which produces deterministic outputs given fixed model weights and inputs.

6 Evaluation and Results

We conduct comprehensive evaluations via automatic metrics, domain experts, layperson, along with LLM-As-Judge for assessing the quality of the model-generated summaries.

6.1 Automatic Evaluation

Evaluation methodology. Automatic evaluation has been performed across three dimensions: *readability*, *content quality*, and *simplicity*.
(i) *Readability*: For measuring text *readability*, we report GFI, FKGL, DC and CLI metrics. Lower values of these metrics indicate simpler text.
(ii) *Content quality*: To measure lexical overlap between reference summary and model output, the standard metrics *ROUGE-1*, *ROUGE-2*, and *ROUGE-L* (Lin, 2004) are used. To capture semantic similarity, *BERTScore* (Zhang et al., 2020) is used. For *ROUGE* and *BERTScore*, we have reported the F1 score.
(ii) *Simplicity*: For *simplicity* measure, *SARI* (Xu et al., 2016) is reported. *SARI* denotes System output Against References and against the Input sentence. It evaluates the quality of content added, deleted, and retained. In this study, *SARI* is used for evaluating how the generated lay summaries are simplified. A higher *SARI* score indicates that the generated lay summary is closer to the reference summary and thus easier for a layperson to comprehend. More details and the significance of *SARI* is discussed in Appendix E.

Evaluation results. As described in Section 3, the test set has two splits: *Test_Complex* (with complex judgements) and *Test_Simpler* (with simpler judgements). For brevity, we report only the results over *Test_Complex* set in the main paper (Table 5). The results of models over *Test_Simpler* split is shown in Appendix F.1 (Table 12).

Model	Method	GFI↓	FKGL↓	DC↓	CLI↓	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	SARI
					Decoder-Only Models					
Llama	ZSI_Direct	9.48	9.04	7.55	9.80	40.14	11.86	17.82	82.58	35.68
	ZSI_Pipeline	9.02	8.80	7.76	10.64	37.15	10.48	16.42	81.57	35.82
	DPO	8.19	7.35	6.78	8.72	37.47	9.10	17.64	81.98	34.55
	SFT	7.27	6.56	6.76	7.80	40.27	12.62	18.20	82.64	36.29
	SFT+DPO	6.75	6.14	6.43	7.69	38.17	9.92	17.23	82.42	35.58
Mistral	ZSI_Direct	11.88	11.17	8.59	11.29	40.17	10.27	17.70	81.81	34.94
	ZSI_Pipeline	11.40	10.66	8.66	11.12	39.05	10.90	17.42	81.67	35.82
	DPO	8.87	7.73	7.21	8.41	33.83	7.04	16.37	82.02	33.71
	SFT	8.03	6.99	7.02	7.55	39.71	10.88	17.91	82.34	35.64
	SFT+DPO	7.63	6.64	6.79	7.85	37.80	9.17	17.48	82.32	35.14
Gemma	ZSI_Direct	10.23	8.85	8.41	9.70	26.52	7.07	13.85	78.56	33.18
	ZSI_Pipeline	10.27	9.97	7.64	10.99	29.84	5.86	14.41	79.46	33.75
	DPO	11.63	9.50	7.90	9.48	23.68	6.81	13.09	82.36	32.24
	SFT	8.11	7.57	7.07	8.22	42.87	14.21	19.37	82.90	36.63
	SFT+DPO	7.45	6.66	6.67	7.55	37.99	9.71	17.35	82.28	35.43
Phi4	ZSI_Direct	9.27	8.49	7.64	9.35	34.53	7.99	15.72	81.32	34.53
	ZSI_Pipeline	9.65	8.61	7.79	8.77	32.29	7.14	14.98	80.47	34.59
	DPO	8.16	6.93	6.87	7.93	32.61	6.84	15.83	81.77	33.71
	SFT	7.48	6.66	6.85	8.06	39.53	10.94	18.57	82.59	35.35
	SFT+DPO	7.32	6.49	6.81	8.01	37.89	9.76	17.56	82.16	35.09
GPT-5*	ZSI_Direct	7.97	7.43	7.85	9.76	51.01	19.67	24.60	84.89	36.78
					Encoder-Decoder Models					
Flan T5	ZSI_Direct	11.78	10.75	8.36	9.43	51.13	25.26	24.83	83.29	37.55
	SFT	8.89	8.92	5.05	8.37	38.88	13.61	20.49	81.54	37.25
Legal Pegasus	ZSI_Direct	11.14	10.99	5.78	9.08	39.80	16.65	21.37	81.93	38.25
	SFT	7.41	7.55	1.39	8.21	28.78	8.01	17.82	79.15	35.39
Legal LED	ZSI_Direct	12.32	11.81	7.65	12.05	27.86	6.73	15.32	77.24	33.89
	SFT	8.59	8.42	6.13	8.32	35.18	11.63	19.90	82.09	36.25

Table 5: Combined readability, quality, and simplification results across models over the *Test_Complex* split of `InLegalLaySum` (*GPT-5 has been evaluated over 200 random samples of the *test* split.). Lower GFI, FKGL, DC and CLI indicate simpler text. For each model, scores marked in *Blue* indicate best readability, *Green* shows best quality, and *red* shows highest simplification across all methods.

Trade-off Between Readability and Semantic Fidelity: For decoder-only models,the SFT method performs the best in terms of summarization quality metrics and SARI, while SFT+DPO performs the best in terms of readability metrics. We observe a consistent drop in scores for quality metrics (*ROUGE*, *BERTScore*) for the hybrid SFT+DPO approach compared to SFT alone, reflecting a trade-off between semantic fidelity and readability. Readability metrics (*FKGL*, *GFI*) primarily capture surface-level properties such as simpler syntax, shorter sentences, and lower structural complexity; hence, they improve after DPO as discussed in Appendix F.4. This trade-off is also aligned with prior observations (Engelmann et al., 2024; Agrawal and Carpuat, 2023). They observed that readability gains do not necessarily align with semantic faithfulness, thereby reducing the scores of overlap-based metrics such as *ROUGE* and semantic similarity measures such as *BERTScore*. Moreover, Xiao et al. (2026) discussed that performing DPO for human preference alignment shifts the model toward preferred outputs (simpler in our case) at the expense of reference matching. Due to the high cost of running GPT-5 on long legal documents, we apply GPT-5 on 200 randomly chosen samples from the *test* split. Comparison of GPT-5 with our best performing open-source model (Llama) over these same 200 samples is reported in Appendix F.2.

Behavior of Encoder-Decoder Models: In the case of Flan T5 and Legal Pegasus, the readability scores improve after doing SFT compared to ZSI, but the quality scores decline because when ZSI was performed with these two models, the generated summaries mostly contain extractive fragments from the judgements. As demonstrated by Datta et al. (2023) extractive summarization yields higher quality scores. As the training dataset contains simplified, abstractive summaries, the summaries' quality scores decrease after performing SFT with Flan T5 and Legal Pegasus.

6.2 Domain Expert Evaluation

Evaluation Setup. We conducted an expert evaluation of the model-generated summaries by the

Model	Mode	Human & LLM Evaluation						Automated Evaluation								
		RC	FC	LTE	Gr	Co	CH	GFI↓	FKGL↓	DC↓	CLI↓	R1↑	R2↑	RL↑	BERT↑	SARI↑
Llama	Human (50)	**3.32**	**3.1**	**3.84**	**3.48**	**3.18**	**3.02**	**6.57**	**5.76**	**6.54**	**7.53**	37.29	10.44	17.36	83.26	35.11
	LLM (50)	**3.18**	**3.22**	**3.94**	**3.26**	**3.16**	**2.98**	**6.57**	**5.76**	**6.54**	**7.53**	37.29	10.44	17.36	83.26	35.11
	LLM (200)	**3.16**	**3.13**	**3.9**	**3.19**	**3.16**	**2.83**	**6.55**	**5.79**	**6.67**	**7.46**	39.88	10.31	17.87	83.04	33.34
Gemma	Human (50)	2.56	2.46	3.1	3.12	2.4	2.46	7.41	6.6	6.89	7.49	39.07	13.34	18.61	**83.42**	**35.99**
	LLM (50)	2.64	2.62	3.38	2.7	2.54	2.5	7.41	6.6	6.89	7.49	39.07	13.34	18.61	**83.42**	**35.99**
	LLM (200)	2.64	2.53	3.45	2.73	2.64	2.45	7.51	6.91	7.07	7.86	43.88	14.55	19.96	83.55	34.46
Flan T5	Human (50)	2.22	2.26	2.2	3	2.24	2.28	10.02	8.99	7.95	8.9	**50.46**	**25.6**	**25.29**	83.43	36.72
	LLM (50)	2.16	2.24	2.36	2.14	2.12	2.08	10.02	8.99	7.95	8.9	**50.46**	**25.6**	**25.29**	83.43	36.72
	LLM (200)	2.19	2.31	2.4	2.21	2.19	2.31	9.57	8.56	8.08	8.96	**56.67**	**32.31**	**30.99**	**84.27**	**35.99**

Table 6: Comparison of model generated lay summaries using human, LLM-As-Judge and automated metrics. RC: Relevance & Coverage, FC: Factual Consistency, LTE: Legal Terms Explained, Gr: Grammaticality, Co: Coherence, CH: Conciseness & Hallucination. Human (50) and LLM (50) shows that evaluation is performed on 50 sampled judgements from the test data; LLM (200) shows LLM-as-a-judge evaluation over 200 random samples of the test data. The highlighted values are the best scores in each mode of evaluation among the different models.

Model	Method	Complex Phrase in Judgement	Simplified Phrase in Model generated Summary
Llama	SFT+DPO	Contingent It was the case of Plaintiff	depends on something happening in the future they argued
Gemma	SFT	collaterals from a whatever source derived	secondary heir income can come from anywhere

Table 7: Expert-identified phrase simplifications in model-generated lay summary

same domain expert who evaluated the *Gemini*-simplified training summaries. We randomly selected 50 samples from the test set to evaluate the outputs of the top 3 models according to the automatic metrics: Llama (SFT+DPO), Gemma (SFT), and Flan T5 (ZSI_Direct). As a result, the expert evaluated 150 model-generated summaries (50 per model).[7] The expert rated each summary on a 5-point Likert scale (1: lowest; 5: highest) according to six mutually discussed metrics (Table 3).

Evaluation Results. Table 6 reports the average scores across all evaluated samples. Although automated metrics indicate that Llama generates simpler summaries and Flan T5 produces higher-quality outputs, the expert preferred Llama summaries overall. The expert also observed that Flan T5 (ZSI_Direct) outputs tend to be largely extractive, retaining many phrases directly from the original judgements; therefore, achieving higher scores in quality metrics.

Table 7 presents examples of complex legal expressions simplified by the models, while Table 15 in Appendix F.4 provides the expert's qualitative remarks. We also observed occasional hallucinations in model outputs, as illustrated by the positive and negative examples in Table 16 (Appendix F.4).

6.3 Layperson Evaluation

Evaluation Setup. Following (Gangavarapu et al., 2025) and (Guo et al., 2026), we recruited two laypersons with no legal background to evaluate the model-generated summaries. Both are proficient in English; one is at a high school level, and the other is a graduate. Each evaluator assessed a total of 60 model-generated summaries, i.e., 30 evaluated outputs for Llama(SFT+DPO) and 30 for Gemma(SFT), resulting in a total of 120 layperson evaluations. We chose these two methods for this evaluation because Llama (SFT+DPO) had the best readability scores and Gemma (SFT) had the best quality scores as per the automatic metrics. Evaluators are asked to rate the model-generated summaries on five dimensions (listed in Table 8) on a Likert scale of 1-5 (1: lowest; 5: highest).[8]

Evaluation Results. Table 9 reports the average scores for the layperson evaluation, providing a direct measure of comprehensibility by layperson beyond automated readability scores. The

[7]This number of evaluated samples is more than in prior legal summarization studies, e.g., 35 summaries (7 algorithms for 5 documents) in (Shukla et al., 2022) and 75 documents (25 documents across 3 datasets) in (Deroy et al., 2024).

[8]Note that different metrics are used for expert and layperson evaluations, as laypersons may not recognize hallucinations or complex legal nuances that experts can.

Metric	Description
Sentence Simplicity (SS)	How easy are the sentences to understand (not too long or complex)?
Term Simplicity (TS)	Are legal domain-specific terms replaced with alternatives or explained in the summary?
Informativeness (In)	Does the summary give an idea of the legal questions/issues involved in the case?
Coherence (Co)	Are ideas presented in a logical and connected manner?
Grammaticality (Gr)	Is the summary grammatically correct?

Table 8: Description of metrics used for layperson evaluation

Level	Model	Method	SS	TS	IN	Co	Gr
High School	Llama	SFT+DPO	4.07	3.93	4.40	3.80	4.13
	Gemma	SFT	2.97	3.30	3.07	2.57	3.10
College Grad.	Llama	SFT+DPO	3.97	3.60	3.57	3.10	3.57
	Gemma	SFT	3.47	3.20	3.47	3.00	3.27

Table 9: Layperson evaluation over 30 samples of model outputs generated by each of Llama(SFT+DPO) and Gemma (SFT); Layperson with high school level seems to give better results compared to graduate level.

layman evaluators gave higher scores to Llama (SFT+DPO), which is consistent with the trends in the expert evaluation (shown in Table 6).

6.4 LLM-As-Judge Evaluation

Evaluation Setup. We also conducted LLM-As-Judge evaluation since large-scale human evaluation is constrained by the time-intensive nature of the task and the limited availability of the domain experts. Following (Viswanathan et al., 2025), we choose the *Qwen-2.5-32B-Instruct*[9] model for the LLM-as-a-Judge evaluation. The prompt used for this evaluation is shown in Appendix B (Figures 14 and 15). First, we performed the LLM-As-Judge evaluation on the same set of model-generated summaries that had been evaluated by the human expert, to compare the scores with those of the human expert. We further conduct LLM-As-Judge evaluation on 200 random samples from the test set.

Evaluation Results. The results of LLM-As-Judge evaluation for Llama (SFT+DPO), Gemma (SFT), and Flan T5 (ZSI) are reported in Table 6. We observe that the average scores from the LLM-As-Judge evaluation are broadly consistent with those from the human evaluation. To further verify the reliability of the LLM-As-Judge evaluation, we compute Pearson and Spearman correlations between expert scores and LLM-As-Judge scores. Detailed results are reported in Appendix F.3 (Table 14). The results show moderate positive correlations between human evaluation and LLM-As-Judge scores, with the highest Pearson and Spearman values of 0.5930 and 0.6279, respectively, for the Relevance & Coverage (RC) metric. The observed moderate correlations are consistent with prior work on LLM-As-Judge evaluation for legal summarization (Liu et al., 2023). Gemma-based evaluation demonstrated the highest agreement with expert scores across most evaluation dimensions. Stronger correlations were achieved for the evaluation metrics – Relevance & Coverage (RC) and Factual Consistency (FC). However, lower agreement on subjective metrics such as Conciseness & Hallucination indicates the limitation of LLM-As-Judge evaluation that aligns with prior work (Evans et al., 2024).

Table 6 also shows the average LLM-As-Judge scores over a larger set of 200 documents from the test set. Llama (SFT+DPO) achieves the highest overall evaluation scores, which is consistent with the human expert evaluation. As moderate correlation alone cannot fully validate the pipeline, which is why we use it only as a supporting signal alongside our primary expert human evaluation, not a replacement for it.

7 Conclusion

We present `InLegalLaySum`, the first benchmark for lay summarization in the Indian legal domain, and conduct a systematic benchmarking with different models and training strategies. Our comprehensive evaluation (via automatic metrics, LLM-As-Judge, domain experts, and laypeople) demonstrates that the SFT+DPO method shows a good balance between semantic preservation and readability improvement, as it consistently outperforms the baseline ZSI_Direct results. Although DPO improves simplicity, it shows a decline in overlap-based quality metrics, reflecting a trade-off between semantic fidelity and simplification. Overall, the SFT+DPO method enhances readability and also amplifies the trade-off, illustrating the alignment tax associated with preference optimization.

[9] https://huggingface.co/Qwen/Qwen2.5-32B-Instruct

Limitations

Although our study provides valuable insights for legal lay summarization, several limitations remain. First, though automatic evaluation, human expert evaluation, and LLM-As-Judge evaluation are performed, the human expert evaluation is limited since summaries of only 100 samples from the train set and 150 samples ($50 \times 3 = 150$) from the test set were evaluated by a single domain expert. It should be noted that this evaluation is time-consuming, and the expert invested substantial time in reading each summary and identifying various types of errors.

Second, readability metrics such as GFI, FKGL, DC, and CLI assess surface-level simplicity but fail to assess factual consistency, which is critical in legal lay summarization. Nevertheless, these metrics have been used to judge the readability of summaries / text in many prior works in both legal and other domains.

Finally, as our experiments focused only on English and on Indian case judgements, the findings may not generalize to other languages or court judgements of other countries.

Ethical Considerations

In this work, we present a dataset for summarizing court judgements to make them more understandable to laypeople. The dataset is made available specifically for academic, non-commercial research, under the *Creative Commons Attribution-NonCommercial-ShareAlike* (CC BY-NC-SA 4.0) license.

The AI methods / models designed in this study are only to provide summaries of legal documents to law professionals or laypeople, and are likely to help common people understand court judgements. These AI methods / models are *not* expected to be integrated directly into the decision-making process of the judicial system.

The expert evaluations of the summaries are conducted by a Law domain expert – an LL.M. (Master of Laws) graduate from a reputed Law school in India – who is a member of our research lab and the third author of the paper. Hence, the annotations were done as part of the domain expert's regular duties. The layperson annotators were informed about the goals of the project and were paid a mutually agreed honorarium.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions. We also thank the annotators who helped us in the layperson evaluation of the summaries.

This research is partially supported by the IIT Mandi iHub and HCi Foundation (iHub), through the project titled "Large Language Model for Legal Assistance".

References

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. SIMSUM: Document-level text simplification via simultaneous summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944, Toronto, Canada. Association for Computational Linguistics.

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302, Singapore. Association for Computational Linguistics.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2024. Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law*, 33:1007 – 1050.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Björn Engelmann, Christin Katharina Kreutz, Fabian Haak, and Philipp Schaer. 2024. ARTS: Assessing readability & text simplicity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14925–14942, Miami, Florida, USA. Association for Computational Linguistics.

Julia Evans, Jennifer D'Souza, and Sören Auer. 2024. Large language models as evaluators for scientific synthesis. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*,

pages 1–22, Vienna, Austria. Association for Computational Linguistics.

Haritha Gangavarapu, Giridhar Kaushik Ramachandran, Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2025. Adapting biomedical abstracts into plain language using large language models. *Preprint*, arXiv:2501.15700.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yue Guo, Jae Ho Sohn, Gondy Leroy, and Trevor Cohen. 2026. Are llm-generated plain language summaries truly understandable? a large-scale crowdsourced evaluation. *Journal of Biomedical Informatics*, 179:105038.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Manuj Malik, Zheng Zhao, Marcio Fonseca, Shrisha Rao, and Shay B. Cohen. 2024. Civilsum: A dataset for abstractive summarization of indian court decisions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2241–2250, New York, NY, USA. Association for Computing Machinery.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Shufan Ming, Yue Guo, and Halil Kilicoglu. 2025. Towards knowledge-guided biomedical lay summarization using large language models. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 285–297, Albuquerque, New Mexico. Association for Computational Linguistics.

Gianluca Moro, Leonardo David Matteo Magnani, and Luca Ragazzi. 2025. Legal lay summarization: exploring methods and data generation with large language models. *Artificial Intelligence Review*, 59.

Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Kumar Bhowmick, Aditya Vempaty, Prasenjit Dey, Ravi Kokku, Pawan Goyal, and Niloy Ganguly. 2025. Introducing spotlight: A novel approach for generating captivating key information from documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35461–35489, Suzhou, China. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document

summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Pawan Trivedi, Digha Jain, Shilpa Gite, Ketan Kotecha, Anant Bhatt, and Nithesh Naik. 2024. Indian legal corpus (ilc): A dataset for a datasetnbsp;summarizing indian legal proceeding using natural language. *Engineered Science*, 27:1022.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. 2025. Checklists are better than reward models for aligning language models. *Preprint*, arXiv:2507.18624.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William K. Cheung, and Chenghua Lin. 2025. Overview of the BioLaySumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 365–377, Viena, Austria. Association for Computational Linguistics.

Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Zongyue Li, Ruirui Lei, Wanggui He, Luu Anh Tuan, Long Chan, Hao Jiang, Zhou Zhao, and Fei Wu. 2026. A survey of direct preference optimization: Datasets, theories, variants, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):13079–13098.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Appendix

A More details of InLegalLaySum dataset

A.1 Data Curation of lay summaries from the BarAndBench source

BarAndBench is a premier legal news website that reports legal news, court proceedings, landmark judgements and developments related to the Indian judiciary. The platform regularly publishes articles covering decisions from the Supreme Court of India, various High Courts and other legal institutions. In addition to reporting legal proceedings, BarAndBench also provides case summaries written in a comparatively simpler manner, making complex legal matters more accessible to non-expert readers. In our work, these expert-written simplified summaries are used as a source for constructing lay summaries corresponding to Indian legal judgments.

To collect the dataset, we utilized the Bing Search API to access the official website of *BarAndBench* and collected the news articles (which we considered as summaries). We select only those articles for which (i) the corresponding judgement link was present (we followed the link to open the PDF containing the judgement and extracted its text) and (ii) the readability scores (GFI and FKGL) for the article are below 12. Finally, the articles meeting these criteria are paired with their respective judgements and these pairs are considered as the (judgement, summary) datapoints.

A.2 Document length distributions in InLegalLaySum

This section presents the distribution of document lengths, i.e., the number of words in the case judgments and summaries of InLegalLaySum. The graphs show the distribution of document lengths across various length ranges. Each point on these graphs represents a bin for a range of word counts in the documents.

Figure 3 shows the document length distribution of Judgements in the *train* split. Similarly, Figures 4 and 5 show the document length distribution of expert-written and *Gemini*-generated simplified summaries in the *train* split, respectively.

Figure 6 shows the document length distribution of Judgments in the *test* split. Similarly, Figure 7 shows the document length distribution of expert-written summaries in the *test* split. The document length distributions for both the train and test sets show a highly imbalanced pattern. Most judgements contain a smaller number of tokens, while only a limited number of cases are extremely long. The distributions also show a long-tail behaviour, where a few outlier documents contain very high token counts compared to the majority of the dataset. Despite minor variations in frequency, both the training and test sets follow a very similar overall distribution pattern.

Metric 1	Metric 2	Pearson	Spearman
GFI	FKGL	**0.9658**	**0.9166**
GFI	DC	0.5305	0.4107
GFI	CLI	0.1728	0.2411
FKGL	DC	0.4145	0.2387
FKGL	CLI	0.2861	0.3877
DC	CLI	0.2141	0.2309

Table 10: Pearson and Spearman correlation between different readability metrics of the expert summaries of the InLegalLaySum dataset.

A.3 Use of GFI and FKGL for creating the test set

Out of the four readability metrics – FKGL, GFI, DC, CLI – we used the FKGL (Flesch-Kincaid Grade Level) and GFI (Gunning Fog Index) metrics for deciding the test split. This choice is explained as follows.

First, the FKGL and GFI readability metrics are widely used to assess readability of various types of texts, including summaries. Prior works like (Xiao et al., 2025) employed FKGL to assess whether generated biomedical lay summaries are readable and accessible to non-expert audiences. Maddela et al. (2021) also applied FKGL to quantitatively evaluate the readability level of simplified text. Second, we measured Pearson and Spearman correlations between the different readability metrics, reported in Table 10. The values clearly show the high correlation between GFI and FKGL compared to correlation among other readability metrics. For these reasons, we selected these two metrics (GFI and FKGL) for creation of the test split of the InLegalLaySum dataset.

A.4 Qualitative Analysis of Training data

Example of an expert written summary and it's simplified summary generated by Gemini are shown in Figure 9 and Figure 10 respectively. The qualitative analysis of the summaries generated by Gemini-2.0-Flash of the training data is shown in Table 11. It highlights that the generated summaries are gen-

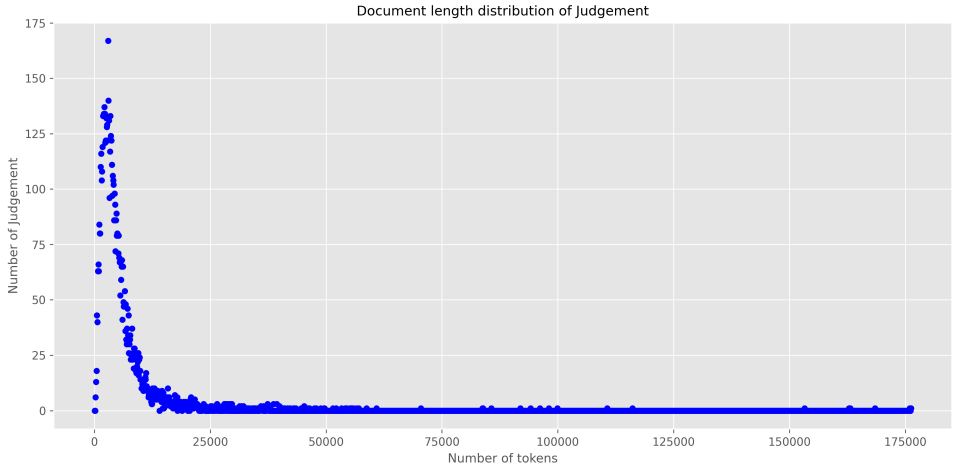


Figure 3: Document length distribution of Judgements in Train set of `InLegalLaySum` dataset

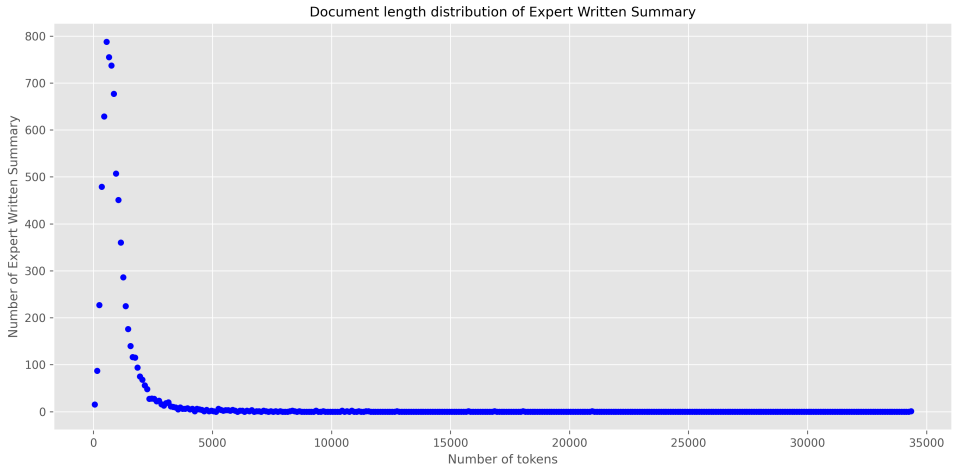


Figure 4: Document length distribution of Expert written summaries in Train set of `InLegalLaySum` dataset

erally effective in simplifying complex legal concepts, improving readability and preserving the core meaning of the expert written summary while making them more accessible to lay readers.

B Prompts

The details of the prompts designed for the experimentation are as follows:

- Figure 8 shows the prompt for generating simplified summaries from expert-written summaries in the *train* set
- Figure 11 represents the prompt used for inference.
- **Prompts for ZSI_Pipeline:** The prompts for the summarization and simplification steps are shown in Figure 12 and Figure 13 respectively.
- **Prompt for evaluation by LLM-As-Judge:** The detailed system prompt and user prompt for LLM-as-judge evaluation are shown in Figure 14 and Figure 15 respectively.

C Models

The detailed list of all models used in the experiments is provided below :

- **Legal-LED-Base-16384**[10] **(Legal LED)**: A 162M parameter encoder-decoder model having a context window of 16K, specifically designed for processing long legal text.
- **Legal Pegasus**[11] **(Legal Pegasus)**: A 560M parameter encoder-decoder model having a context window of 1024, adapted from the base Pegasus model, meant for abstractive summarization of legal documents.
- **Flan-T5-Large**[12] **(Flan T5)**: A 780M parameter

[10] https://huggingface.co/nsi319/legal-led-base-16384

[11] https://huggingface.co/nsi319/legal-pegasus

[12] https://huggingface.co/google/flan-t5-large

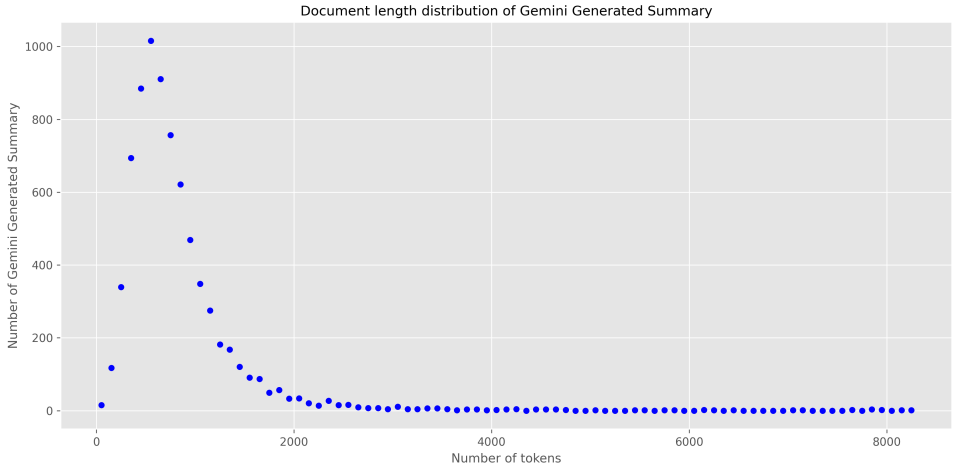


Figure 5: Document length distribution of summaries simplified by Gemini, from the train set of `InLegalLaySum`

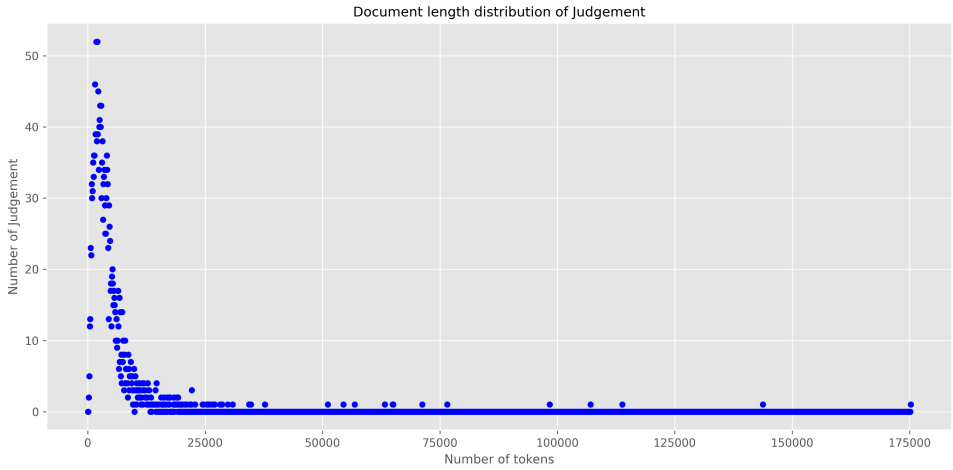


Figure 6: Document length distribution of Judgements in Test set of `InLegalLaySum` dataset

instruction-tuned encoder-decoder model having a context window of 512 designed for diverse NLP tasks.

- **Gemma-2B-IT**(Team et al., 2024)[13] **(Gemma)**: A 2B parameter instruction-tuned decoder-only LLM having a context window of 8K.
- **Llama-3.2-3B-Instruct**[14] (Grattafiori et al., 2024) **(Llama)**: A 3B parameter instruction-tuned decoder-only LLM having a context window of 128K designed for efficient conversational text generation.
- **Phi-4-mini-instruct**[15] **(Phi4)**: A 4B parameter instruction-tuned decoder-only model having a context window of 128K.
- **Mistral-7B-Instruct-v0.3**[16](Jiang et al., 2023) **(Mistral)**: A 7B parameter instruction-tuned decoder-only model having a context window of 32K.
- **GPT-5**[17] **(GPT5)**: An instruction-tuned decoder-only proprietary SOTA LLM for advanced text generation.

D Experimental Setup

Chunking of long documents. To deal with lengthy legal documents, we chunked the documents into m small chunks, where the size of each chunk is the maximum number of tokens (say, n) that the model is designed to accept without truncating (e.g., n = 512 for Flan T5). Then, we asked

[13] https://huggingface.co/google/gemma-2b-it

[14] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[15] https://huggingface.co/microsoft/Phi-4-mini-instruct

[16] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[17] https://developers.openai.com/api/docs/models/gpt-5

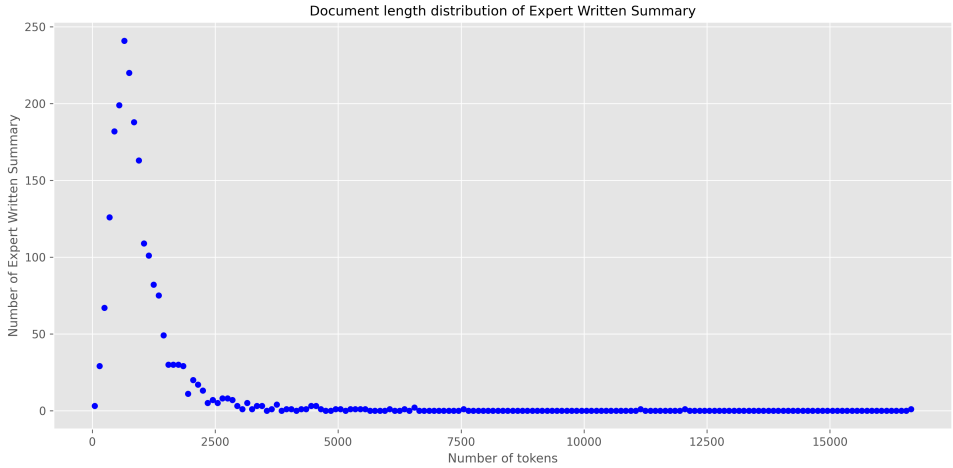


Figure 7: Document length distribution of Expert written summaries in Test set of InLegalLaySum dataset

Human Evaluation Comment	**Observation on Data Quality**
"The summary simplified the contents of VIIth Schedule ... in an easy to understand manner."	Simplification is done without going in depth
"Arbitrator ... someone who settles disputes outside courts."	Complex terms are explained
"Para is written as section."	Common terms are used for accessibility of lay person
"Although, the term retrospectively is not explicitly explain, it's clear from one of the sentences."	Simplified summaries generally retain contextual understanding

Table 11: Qualitative Analysis of Training data

the model to provide a summary of k tokens for each chunk, and append the chunk-wise summaries in the same order in which the chunks appear in the document, such that the combined summary (of total mxk tokens) is almost equal in length to the reference summary. Models with long context windows (e.g., 16,384 tokens for Legal LED) successfully handled almost all the case judgments in InLegalLaySum without truncation or chunking.

Individual gold reference summary lengths are not used at inference - no sample's output length is conditioned on that sample's own reference summary. The max_length parameter is a single fixed value applied uniformly to every test sample. It was set once, based on the average reference summary length across the whole dataset, and is not adjusted per sample. This is why we describe the combined summary length as "almost equal to" rather than "equal to" the reference length - it reflects a corpus-level average, not per-sample matching.

Chunk size during input pre-processing is a separate, unrelated choice, determined solely by each model's maximum input token limit - this applies only to FlanT5 and Legal Pegasus, which have limited input context windows. So the output length cap is a fixed value derived from aggregate test-set statistics, not from per-sample reference lengths, to avoid any implication of instance-level leakage.

E The SARI metric

SARI was designed specifically to evaluate simplification behavior by comparing the add / delete / keep n-gram operations between the source, system output, and reference, rather than raw content overlap as in ROUGE. This makes it well-suited to our approach, where the goal is not just to compress a legal document but to render it in more accessible language for a lay audience – a transformation ROUGE and BLEU are not designed to capture. This usage is consistent with prior work applying SARI to long-document abstractive lay summarization in adjacent domains, including legal lay summarization (Moro et al., 2025), paragraph-level medical text simplification (Devaraj et al., 2021), and document-level text simplification via simultaneous summarization (Blinova et al., 2023), all of which adopt SARI as the standard metric for this task. We report it alongside other readability and quality metrics to jointly capture both content fidelity and language accessibility, providing a more comprehensive picture of summary quality than either metric alone.

F Results & Discussion

F.1 Result over *Test_Simpler* split

The combined readability, quality and simplification results across models over the *Test_Simpler* split is shown in Table 12. Similar to the trends observed in the *Test_Complex* split (which are reported in the main paper), SFT+DPO generally achieves the best readability scores across most decoder-only models, while SFT consistently obtains better quality scores in terms of ROUGE and Bert scores. Among encoder-decoder models, Flan T5 and Legal Pegasus continue to show high ROUGE, Bert and SARI scores with ZSI_Direct, whereas their SFT variants generate comparatively simpler and more readable summaries. Overall, the trends on the *Test_Simpler* split remain largely consistent with those observed on the *Test_Complex* split, suggesting that the relative strengths and weaknesses of different training strategies remain stable across varying difficulty levels of legal documents.

F.2 Comparison with GPT-5

Running proprietary LLMs such as GPT on large datasets is expensive due to API costs, especially for long-document summarization tasks that involve many inputs and response tokens; hence, we apply GPT-5 to 200 randomly selected samples from the *test* split. To make the GPT-5 results comparable, we now report the readability, quality, and simplification scores for Llama (ZSI_Direct; ZSI_Pipeline; DPO; SFT; SFT+DPO) over the same 200 test samples (for which GPT-5 was benchmarked) in Table 13. Llama model-generated outputs are chosen for comparison with GPT-5 results because Llama is the best-performing model in terms of readability. We observe that Llama (SFT+DPO) achieves better readability scores than GPT-5 (ZSI_Direct) on the same 200 test-set samples.

F.3 Correlation between human and LLM-As-Judge evaluation

Table 14 shows the Pearson and Spearman correlation between human evaluation and LLM-As-Judge evaluation on 50 random samples of test data. Higher correlation scores are observed for Relevance & Coverage and Factual Consistency across most models, indicating that LLM-based evaluators are relatively reliable in assessing the overall content quality and semantic alignment of lay summaries. On the other hand, lower correlation scores are obtained for Grammaticality and Conciseness & Hallucination. This is likely because grammatical issues in generated summaries are often context-dependent, while identifying hallucinated or unnecessarily verbose content in legal summaries requires deeper domain understanding and careful fact verification, where human evaluators still perform better than automated evaluators.

F.4 Analysis

Qualitative remark over some model outputs is shown in Table 15. The table shows that Llama (SFT+DPO) generates comparatively better simplified summaries, but still misses some important facts and explanations of legal terms. Gemma (SFT) produces simpler language in some places, but contains grammatical errors, factual mistakes and hallucinated information. Flan T5 (ZSI) mainly copies text directly from the original judgment, making the summary repetitive, incomplete and less useful for lay users. Overall, the analysis shows that current models still struggle to generate summaries that are both factually accurate and easy for non-experts to understand.

Table 16 shows some examples of positive and negative observations in model outputs of Llama (SFT+DPO), Gemma (SFT) and Flan T5 (ZSI_Direct) identified by domain expert. It is observed that both Llama (SFT+DPO) and Gemma (SFT) demonstrate effective simplification of complex legal content, whereas such simplification is not observed in FlanT5 (ZSI_Direct). Across the models, the expert also identified instances of incorrect claims and hallucinated numerical figures in some generated outputs as part of negative observations.

Model	Method	GFI↓	FKGL↓	DC↓	CLI↓	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	SARI
					Decoder-Only Models					
Llama	ZSI_Direct	9.03	8.46	7.50	9.64	40.71	11.63	17.63	82.59	34.53
	ZSI_Pipeline	8.68	8.30	7.72	10.27	37.85	10.12	16.47	81.70	34.63
	DPO	8.18	7.18	7.05	8.81	38.64	9.27	18.08	82.34	33.24
	SFT	6.89	6.08	6.79	7.56	41.03	12.80	18.42	82.88	35.21
	SFT+DPO	6.57	5.79	6.53	7.47	39.46	10.38	17.56	82.64	34.65
Mistral	ZSI_Direct	11.40	10.59	8.55	11.13	41.16	10.59	17.89	81.95	33.91
	ZSI_Pipeline	11.02	10.26	8.66	11.12	39.96	10.74	17.51	81.74	34.68
	DPO	8.66	7.29	7.44	8.32	33.22	6.89	16.28	82.35	32.15
	SFT	7.59	6.47	7.11	7.65	41.27	11.48	18.49	82.75	34.66
	SFT+DPO	7.49	6.41	6.83	7.56	39.44	9.70	17.87	82.56	34.14
Gemma	ZSI_Direct	9.74	8.24	8.29	9.59	27.74	7.18	14.24	78.87	31.98
	ZSI_Pipeline	10.37	10.13	7.66	11.23	30.68	5.99	14.48	79.67	32.71
	DPO	11.21	9.32	8.13	9.29	21.29	6.17	12.40	82.59	30.45
	SFT	7.73	7.12	7.01	8.02	43.13	14.21	19.37	83.01	35.58
	SFT+DPO	7.23	6.38	6.67	7.46	39.10	10.06	17.64	82.45	34.45
Phi4	ZSI_Direct	8.89	8.05	7.49	9.13	36.22	8.70	16.30	81.53	33.61
	ZSI_Pipeline	9.12	8.18	7.66	8.89	33.23	7.19	15.26	80.66	33.46
	DPO	7.72	6.36	7.09	7.64	31.86	6.40	15.60	82.02	32.16
	SFT	7.15	6.23	6.99	7.97	39.91	11.08	18.84	82.91	34.06
	SFT+DPO	7.14	6.12	6.91	7.85	38.38	9.85	17.88	82.38	33.89
					Encoder-Decoder Models					
Flan T5	ZSI_Direct	9.67	8.78	8.08	9.19	52.34	27.27	26.97	83.56	36.99
	SFT	8.89	8.89	4.08	8.35	36.89	11.97	20.18	80.97	35.82
Legal Pegasus	ZSI_Direct	10.05	10.02	5.49	9.08	39.71	15.87	21.95	81.70	37.03
	SFT	7.09	7.29	1.33	8.29	28.18	7.31	18.48	78.94	33.81
Legal LED	ZSI_Direct	11.96	11.43	7.62	12.05	29.44	7.21	15.96	77.40	32.72
	SFT	7.68	7.36	6.16	8.25	36.72	11.85	20.61	82.37	35.29

Table 12: Combined readability, quality, and simplification results across models over *Test_Simpler* split. Lower GFI, FKGL, DC and CLI indicate simpler text. Scores marked in *Blue* indicate best readability, *Green* indicate best quality and *Red* indicate highest simplification performance.

Models	Method	Readability				Quality				Simplification
		GFI↓	FKGL↓	DC↓	CLI↓	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	SARI
Llama	ZSI_Direct	9.22	8.65	7.53	9.32	41.02	11.93	17.91	82.78	34.07
	ZSI_Pipeline	8.84	8.49	7.83	10.22	38.56	10.61	16.71	81.87	34.31
	DPO	8.15	7.28	6.97	8.86	38.68	9.33	18.22	82.45	32.74
	SFT	6.97	6.14	6.77	7.45	41.3	12.57	18.41	82.82	34.48
	SFT+DPO	**6.57**	**5.77**	**6.56**	**7.52**	40.07	10.4	17.81	82.76	34.15
GPT5	ZSI_Direct	7.97	7.43	7.85	9.76	**51.01**	**19.67**	**24.6**	**84.89**	**36.78**

Table 13: Performance comparison of Llama and GPT5 over methods across readability, quality and simplification metrics.

Model	Method	Metric	Pearson	p-value	Spearman	p-value
Llama	SFT+DPO	RC	0.4069	0.0034	0.4562	0.0008
		FC	**0.5884**	0.0001	**0.6246**	0.0001
		LTE	0.4207	0.0023	0.4349	0.0016
		Gr	**0.4880**	0.0003	**0.4773**	0.0005
		Co	0.3296	0.0194	0.3412	0.0153
		CH	0.3501	0.0127	**0.3632**	0.0095
Gemma	SFT	RC	**0.5930**	0.0001	**0.6279**	0.0001
		FC	0.5190	0.0001	0.5784	0.0010
		LTE	**0.4450**	0.0012	**0.5306**	0.0001
		Gr	0.3979	0.0042	0.3627	0.0096
		Co	**0.4615**	0.0007	**0.4869**	0.0003
		CH	0.3231	0.0221	0.3564	0.0111
Flan T5	ZSI	RC	0.4519	0.0009	0.4817	0.0004
		FC	0.4505	0.0010	0.4486	0.0011
		LTE	0.4132	0.0029	0.4339	0.0016
		Gr	0.3084	0.0293	0.2349	0.1006
		Co	0.3173	0.0247	0.3533	0.0118
		CH	**0.3848**	0.0058	0.3408	0.0154

Table 14: Pearson and Spearman correlation analysis between human evaluation scores and LLM-based evaluation scores across different evaluation metrics. RC: Relevance & Coverage, FC: Factual Consistency, LTE: Legal Terms Explained, Gr: Grammaticality, Co: Coherence, CH: Conciseness & Hallucination; The highlighted scores indicate the highest correlation among each evaluation metrics.

```
You are a legal expert skilled in simplifying complex legal summaries. Rewrite
the following layman summary so it is easy for a high school student (11th–12th
grade level) to understand.
Your task is to:
- Use simple vocabulary and short, clear sentences.
- Preserve all original structure, including paragraph breaks and sections if
present.
- Keep each paragraph logically consistent and aligned with the original idea.
- Clarify or explain legal terms using plain English, only when necessary.
- Ensure the rewritten version is easy to read, with a Flesch-Kincaid Grade Level
below 10.
Do NOT:
- Add FKGL scores
- Add section headers unless already present
- Use bullet points
- Add introductions, summaries, or personal opinions
Layman Summary: {Insert the layman summary here}
Return only the rewritten version with the original structure preserved.
```

Figure 8: Expert written summaries of Train set of `InLegalLaySum` dataset are simplified by *Gemini-2.0-Flash* with this prompt for creating the corresponding simplified summaries of Judgement in the Train set

Part of Lay Summary	Human (Domain expert) Remarks
Llama (SFT+DPO)	
Indian Law Case The company being taxed had insurance for its buildings and equipment. This insurance covered losses of money made from selling things, the cost of keeping the buildings, and commissions paid to agents. The company's buildings were destroyed in a fire on January 18, 1844. ..... The idea that this money isn't income because it's only paid if there's a loss, and that it's more like a gift from someone else, isn't correct. Also, the idea that the person doesn't earn money from the work they do, and therefore doesn't have any income, isn't correct. This also isn't true if there is no income or profit.	**Incorrect**: The year when fire occurred, amount paid by insurance companies, money earned from this type is considered taxable, not an exemption.. **Simplification**: "Consequential loss" and Section (4)(3)(vii) of the Act are simplified. **Factual Consistency**: The facts of the case are correct, except for the few missed and incorrect points. Few cited cases in the summary are also not correct. **Explanation of legal terms**: Since it's a case about income-tax, few terms like assessment year, financial year, assessee, etc. should have been explained.
Gemma (SFT)	
The company was insured against losses in profits, standing and agency commissions. The fire broke out on January 1, 4, 94. The insurance companies paid the company Rs.14 lakhs. The income was treated as income and the company was taxed accordingly,.....The Court agreed with the High Courts that the Company paid income tax. [8] The appeal was dismissed. [9] The case of Commissioner of Income Tax vs William E. Wallace & CO. was followed. [This case was followed] 1. The taxes are only paid after certain steps have been taken. 7. The business must follow rules before the money can be paid out. 9. The high court was correct in saying that the 2.	**Incorrect grammar** in the first paragraph-income was income; income was treated as income. **Incorrect**: One way is by buying and selling insurance against loss in profit. Shaw Wallace & co vs The Commissioner of Inland Revenue. In the last paragraph, the argument of appellant is treated as judgment of court, which is incorrect. **Hallucination**: The company paid Rs.24 lakhs in insurance; company paid tax on the Rs 240,00000 paid in insurance payments. **Simplification**: "from a whatever source derived" has been simplified as "income can come from anywhere". Meaning of the word "purport" has not been explained.
Flan T5 (ZSI)	
Appeal No. 55 of 1950. Appeal by special leave from the Judgment and Order dated March 18. 1949, of the High Court of Judicature at Bombay (Chagla C. J. 178 and Ten dolkar J.) in Income tax Reference 14,00,000 for loss of Profits, Standing Charges and Agency Commission of the above Co. 's Mills, ... The learned Solicitor General relies strongly a clause which appears in three of the four policies with which we are concerned. That is a clause which states that the insured must do all he can to minimise the loss in profits and until.	**Irrelevant, repetitive and incomplete information**: Text in the beginning of summary; 2nd paragraph; excerpts from the original judgment in verbatim. **Verbatim information**: The summary is excerpt from the original judgment. No explanation of **legal terminology**.

Table 15: Comparison of remarks by law domain expert on outputs generated by Llama, Gemma and Flan T5 for the same sample case.

Paradigm	Positive Example	Negative Example
Llama (SFT+DPO)	Simplifies "consequential loss" and Sec.~4(3)(vii) accurately, preserving legal meaning.	Incorrect claim on taxability of insurance payout; miscited case references; unexplained terms (assessee, assessment year).
Gemma (SFT)	"From whatever source derived" -> "income can come from anywhere" - clear, accurate paraphrase.	Hallucinated figures (Rs.~24 lakhs payout); appellant's argument misrepresented as court's judgment.
FlanT5 (ZSI)	No meaningful simplification observed.	Verbatim reproduction of judgment text; irrelevant procedural boilerplate; no legal terms explained.

Table 16: Detailed analysis of model outputs showing both positive and negative examples performed by domain expert.

In a case where death is due to injuries or wounds caused by a lethal weapon, it has always been considered to be the duty of the prosecution to prove by expert evidence that it was likely or at least possible for the injuries to have been caused with the weapon with which, and in the manner in which, they are have been caused. Where in a case of murder, the prosecution case was that the accused shot the deceased with a gun, but it appeared likely that the injuries on the deceased were inflicted by a rifle and there was no evidence of a duly qualified expert to prove that the injuries were caused by a gun, and the nature of the inajuries was also such that the shots must have been fired by more than one person and not by one person only, and the prosection had no evidence to show that another person also shot, and the High Court, though realis ing that there was thus a gap in the prosecution evidence, convicted the accused placing reliance on the oral evidence of 3 witnesses which was not, disinterested: Held, that, the present case fell within the rule laid down in Pritam Singh vs The State ([1959] S C R. 453) inas much as the appellant had been convicted notwithstanding the fact that evidence was wanting on a most material part of the prosecution case, and the conviction could not therefore be upheld, 822 Held also, that the standard of proof which is required in regard to the plea of alibi must be the same as the standard which is applied to the prosecution evidence and in both cases it should be a reasonable standard.

Figure 9: Example of an expert-written summary of a judgement in the Train set of InLegalLaySum dataset

If someone dies from injuries caused by a deadly weapon, the people bringing the case (the prosecution) must show that the weapon could have caused the injuries in the way they claim. They usually need an expert to prove this. For example, in a murder case, the prosecution said the accused person shot the victim with a gun. But it looked like the injuries were from a rifle. There was no expert to say the gun could have caused the injuries. Also, the injuries suggested more than one person fired shots, but the prosecution only accused one person. The High Court knew the prosecution's evidence was lacking but still found the accused guilty based on witness statements. The Supreme Court decided this was wrong. The case was like one they had seen before (Pritam Singh vs The State). The accused person was found guilty even though the prosecution didn't prove a key part of their case. So, the conviction was overturned. The Supreme Court also said that when someone claims they were somewhere else (an alibi), the proof needed is the same as for the prosecution's case. Both sides need to provide a reasonable level of proof.

Figure 10: Example of a Gemini-generated simplified summary in the Train set of InLegalLaySum dataset – the simplified version of the summary in Figure 9.

SYSTEM PROMPT

You are an expert in Indian law and simplifying legal content. Your task is to summarize complex Indian legal documents so that a reader with up to a high school education can easily understand the summary. Avoid domain-specific complex legal terms and when necessary, provide simple explanations of legal terms in parentheses next to the specific term. Keep the response sentences short and simple. Respond only in English.

USER PROMPT

Please write a simplified summary for the following Indian legal judgment so that the summary is suitable for readers up to high school(11th - 12th grade level) education. Please ensure the summary retains the key facts, core legal issues, arguments and the final decision of the court, without adding any new information or altering the meaning. Do NOT include any closing remarks, disclaimers, notes or offers for further assistance, section headers, personal opinion. Do not include any closing remark or offers for further help. Respond only in English.
Judgment: <JUDGMENT>
Simplified Summary:

Figure 11: Prompt used for inference in Zero Shot Inference (ZSI_Direct)

SYSTEM PROMPT

You are an expert in Indian law. Your task is to summarize complex Indian legal documents. Respond only in English.

USER PROMPT

Please write a summary for the following Indian legal judgment. Please ensure the summary retains the key facts, core legal issues, arguments and the final decision of the court, without adding any new information or altering the meaning. Do not include any closing remarks or disclaimers. Avoid providing offers for further assistance. Neither add any section headers nor give any personal opinion. Do not include any closing remark. Respond only in English.
Judgment: <JUDGMENT>
Summary:

Figure 12: Prompt used for summarization in ZSI_Pipeline

SYSTEM PROMPT

```
You are a legal expert skilled in simplifying complex legal summaries.
```

USER PROMPT

```
Rewrite the following summary so it is easy for a high school student (11th–12th
grade level) to understand.
Your task is to:
- Use simple vocabulary and short, clear sentences.
- Preserve all original structure, including paragraph breaks and sections if
present.
- Keep each paragraph logically consistent and aligned with the original idea.
- Clarify or explain legal terms using plain English, only when necessary.
- Ensure the rewritten version is easy to read.
Do NOT:
- Add section headers unless already present.
- Use bullet points.
- Add introductions, summaries, or personal opinions.
### Summary: <SUMMARY>
### Simplified Summary:
```

Figure 13: Prompt used for simplification in ZSI_Pipeline

SYSTEM PROMPT

You will be given a model generated lay summary along with a complex judgement. A lay summary is referred to as a summary of a complex judgement, which is comprehensible and easier to understand for a reader with up to a high school(12th grade) education.

Your task is to rate the model generated lay summary on some metrics. Please make sure you read and understand these instructions carefully. Please keep these documents open while reviewing and refer to it as needed.

Evaluation Criteria:

1) Relevance & Coverage (RC) : Does the summary cover most of the points from the judgement without missing key information?
- A score of 1 means completely irrelevant or missing major points.
- A score of 2 means covers very few relevant aspects.
- A score of 3 means covers some points but misses several key details.
- A score of 4 means covers most key points with minor omissions.
- A score of 5 means fully captures important aspects of the judgement.

2) Factual Consistency (FC) : Is the model generated lay summary free from hallucination or false claims?
- A score of 1 means major false information and contradiction in model generated lay summary.
- A score of 2 means presence of multiple incorrect or misleading statements.
- A score of 3 means there are some inaccuracies but the core idea is partly correct.
- A score of 4 means mostly accurate with presence of minor errors.
- A score of 5 means completely accurate and no hallucinations.

3) Legal Terms Explained (LTE) : Are legal terms either avoided or clearly explained in simple and easier language?
- A score of 1 means heavy use of unexplained legal jargon.
- A score of 2 means frequent legal terms with little explanation.
- A score of 3 means some legal terms, partially explained.
- A score of 4 means minimal legal jargon, mostly explained.
- A score of 5 means no legal terms or all legal terms clearly explained in simple terms.

4) Grammaticality (Gr) : Is the summary grammatically correct and well formed?
- A score of 1 means very poor grammar, hard to read.
- A score of 2 means frequent grammatical errors.
- A score of 3 means some noticeable errors.
- A score of 4 means minor grammatical issues.
- A score of 5 means grammatically correct and fluent.

5) Coherence (Co) : Do ideas connect logically across sentences in model generated lay summary?
- A score of 1 means completely disjointed.
- A score of 2 means poor flow and hard to follow.
- A score of 3 means some logical flow but inconsistent.
- A score of 4 means mostly coherent with minor issues.
- A score of 5 means fully logical and well-structured.

6) Conciseness & Hallucination (CH) : Is the summary concise while avoiding unnecessary details or fabricated content?
- A score of 1 means contains hallucinations and very verbose.
- A score of 2 means too long with significant incorrect content.
- A score of 3 means some unnecessary details or minor hallucinations.
- A score of 4 means mostly concise with minimal extra content.
- A score of 5 means concise, precise and no hallucination.

Evaluation Steps:

1) Read the complex judgement and model generated lay summary.
2) Rate the model generated lay summary on a scale of 1 to 5 according to the evaluation criteria above.
3) Provide a brief explanation of your rating, referring to specific aspects of model generated lay summary and complex judgement.

Figure 14: System Prompt used for evaluation by LLM-As-Judge on the test set

USER PROMPT

You will be given a model generated lay summary along with a complex judgement. Your task is to rate the model generated lay summary on some metrics. Please make sure you read and understand these instructions carefully. Please keep these documents open while reviewing and refer to it as needed.

COMPLEX JUDGEMENT:
judgement

MODEL-GENERATED LAY SUMMARY:
laysummary

Evaluation Steps:

1) Read the complex judgement and model generated lay summary.
2) Rate the model generated lay summary on a scale of 1 to 5 according to the evaluation criteria above.
3) Provide a brief explanation of your rating, referring to specific aspects of model generated lay summary and complex judgement.

Relevance & Coverage (RC): <1-5>
Factual Consistency (FC) Score: <1-5>
Legal Terms Explained (LTE) Score: <1-5>
Grammaticality (Gr) Score: <1-5>
Coherence (Co) Score: <1-5>
Conciseness & Hallucination (CH) Score: <1-5>

Figure 15: User Prompt used for evaluation by LLM-As-Judge on the test set