“Decode the Law": Towards Legal Text Simplification with Large Language Models

Mohammed Danish Rabbani[1], Subhadeep Roy[2], Sayantan Mitra[3], Tulika Saha[1]
[1]International Institute of Information Technology Bangalore, India
[2]University of Technology Nuremberg, Germany
[3]Samsung Research Institute Bangalore, India
{mohammed.danishrabbani, tulika.saha}@iiitb.ac.in

Abstract

Legal documents are often verbose and structurally complex, posing significant barriers to public understanding and equitable access to justice. Despite growing interest in text simplification, efforts targeting the legal domain remain limited by a lack of robust, high-quality resources. In this paper, we address this gap by introducing *SIMPLE-LAW*, a curated benchmark dataset of over 6,000 aligned pairs of original and simplified legal passages, specifically constructed to facilitate research in legal text simplification by leveraging large language models (LLMs). We evaluate this dataset across both in-context learning and parameter-efficient fine-tuning paradigms using a range of state-of-the-art LLMs, with Unsloth variants of Mistral, LLaMA-3.2, Gemma, and Qwen-2.5. We assess performance using BERTScore, ROUGE, SARI, and a hallucination detection score, to capture both fidelity and readability. Results show that fine-tuned models significantly outperform in-context learners in terms of simplification quality and factual consistency. By offering a new dataset, rigorous evaluation, and baseline comparisons, our work provides a critical foundation for developing transparent and accessible AI systems in the legal domain.

Keywords: Legal Text, Simplification, Legal Domain, Large Language Models

1. Introduction

Legal texts are fairly difficult to comprehend due to their formal structure, domain-specific jargons, and complex sentence formations. This complexity not only impedes access to justice (Gibbons, 2003) but also undermines transparency and informed decision-making amongst the general populace. Legal systems globally assume that individuals understand their rights and obligations, yet legal documents such as contracts and court orders are often inaccessible to the general public, especially the non-legal laypersons. This disproportionately affects vulnerable groups, as noted by the UNDP (United Nations Development Programme, 2022). To counter this, initiatives like the European Commission’s Clear Writing campaign (European Commission, n.d.) and India’s efforts to simplify legal language (PWOnlyIAS, 2025) aim to improve legal accessibility. As digital governance and cross-border legal services grow, the need for comprehensible legal content is becoming a global priority, particularly in low and middle-income countries where barriers are both linguistic and socio-economic. Legal simplification can empower under-served populations by facilitating legal compliance and reducing dependency on costly legal assistance.

Despite significant progress in natural language processing (NLP) for text simplification (Alva-Manchego et al., 2020a), domain-specific resources for the legal field remain scarce. Legal texts pose unique challenges, such as maintaining legal precision while simplifying complexity (Garimella et al., 2022). To address this gap, we present *SIMPLE-LAW*, a curated dataset for legal text simplification utilizing Large Language Models (LLMs). We evaluate Unsloth-optimized versions of Mistral (Unsloth, 2025c; AI, 2023), LLaMA-3.2 (Unsloth, 2025a,b; Grattafiori and et al., 2024), Gemma (Unsloth, 2024; DeepMind, 2024), and Qwen-2.5 (Unsloth, 2025d,e; Team, 2024), through both in-context and parameter-efficient fine-tuning approaches on this dataset. Our comprehensive evaluation, using automated and human metrics, establishes a benchmark for legal text simplification and supports future research in this emerging area.

Research Questions. We address the following research questions in this paper: **(i)** Can industry-grade LLMs be used to curate domain-specific datasets? **(ii)** How effective is the use of semi-synthetic datasets in mimicking the real-world legal simplification task? **(iii)** How do fine-tuned LLMs perform compared to the in-context learning strategy for the legal simplification task?

Key Contributions. The key contributions of this paper are as follows: **(i)** We introduce a semi-synthetic dataset, *SIMPLE-LAW*, exploiting the power of LLMs to address the limitation of existing datasets; **(ii)** The *SIMPLE-LAW* dataset has been evaluated using human evaluation and various other automated techniques to quantify the quality of the curated dataset; **(iii)** We benchmark *SIMPLE-LAW* on a wide range of general-purpose

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 631–641
11–16 May 2026.

LLMs by experimenting with both fine-tuning and in-context learning strategies.

2. Related Work

In this section, we discuss some impactful work relevant to this field of research.

Text Simplification. Early approaches to text simplification are well-surveyed in (Siddharthan, 2014), covering rule-based and statistical techniques across lexical, syntactic, and discourse levels. A broader review in (Al-Thanyyan and Azmi, 2021) critiques existing simplification datasets like WikiLarge and Newsela for their domain mismatch with legal texts. The ACCESS model introduced by (Martin et al., 2020) demonstrates superior controllable simplification, achieving a SARI score of 41.87 on WikiLarge. Prompt-based controllability via LLMs is explored in (Madaan and et al., 2022), enabling human-in-the-loop interventions. Reinforcement learning approaches with readability-focused reward functions are proposed in (Nishihara and et al., 2022). Notably, (Xu et al., 2015) introduced a widely-used large-scale dataset aligned across grade levels, while (Devaraj and et al., 2023) released a medical domain-specific simplification corpus.

Legal Text Simplification. (Garimella et al., 2022) provided in-depth insight into legal domain challenges, notably the lack of complex to simple data available and legal jargon that complicates the comprehension of lay people. (Cemri et al., 2022) shows unsupervised approaches such as USLT leverages domain specific LMs like Legal-BERT for lexical and sentence-level simplification, demonstrating initial gains but still limited by data scarcity. However, Legal-BERT support only 512 input tokens which is not sufficient for this task. (Gallegos, 2022) has shown the usage of summarization & simplification together produced better results by fine-tuning and concluded with a call for higher-quality legal datasets to improve LLMs for the law domain. (Athugodage et al., 2024) introduced the first large-scale parallel corpus for legal text simplification in Russian language, derived from the *Rossiyskaya Gazeta* legal papers. Their study systematically compared transfer learning approaches using T5 and GPT architectures, evaluating performance using both automated metrics (ROUGE, SARI, BERTScore) and readability indices (Flesch-Kinkaid, Gunning Fog). Their work highlighted the superior performance of fine-tuned GPT models on legally simplified corpora, reinforcing the potential of domain-adapted LLMs in improving accessibility of complex legal language. Recently, (Ujwal et al., 2024) introduced a dataset for Long-form Question Answer dataset in the legal domain leveraging the power of LLMs.

3. Dataset

As established earlier, legal text simplification suffers from the absence of a standardized, large-scale parallel corpus across varied legal genres, limiting consistent evaluation. This section introduces our curated dataset, *SIMPLE-LAW*[1].

3.1. Data Collection & Source

The legal texts were obtained from a diverse range of public legal document repositories and legal information portals spanning multiple jurisdictions, including India, the United States (US) and the United Kingdom (UK). The primary goal behind this multi-jurisdictional selection was to ensure diversity in legal language, syntactic complexity, and jurisprudential traditions, thereby strengthening the generalizability and robustness of our legal text simplification framework. We curated a total of 6,230 legal text samples, consisting of complex legal laws or statements. These samples were distributed approximately as follows: (i) Indian sections and acts: approximately 3000+ samples sourced from (Lighteval, 2023), https://www.indiacode.nic.in/ etc.; (ii) USA Laws: 1500+ samples sourced from https://www.usa.gov/laws-and-regulations, https://law.justia.com/ etc.; and (iii) UK Laws: 1500+ samples sourced from https://www.legislation.gov.uk/. These sources include judicial opinions, statutory provisions, legal notices, and contractual clauses ensuring a broad representation of legal language patterns and terminologies. All data samples were collected from publicly accessible resources and were pre-processed to improve formatting.

3.2. Data Generation

Next, the focus was to create the simplified versions of these legal texts. A semi-supervised approach was leveraged which employed the GPT-3.5-Turbo model (OpenAI, 2023) via OpenAI's API. A current version of the GPT-series, i.e., GPT-4 was not taken into consideration due to resource constraints. We utilized a mixture of 1-shot, 2-shot, and chain-of-thought (CoT) prompting techniques to generate simplifications of the legal texts ensuring quality and coherence. The examples used in these in-context strategies were gold-standard raw text-simplification pairs curated in collaboration with two legal experts from the author's affiliation to ensure that the prompts modeled high-quality simplification patterns rooted in domain expertise.

[1]The dataset & codes are availale at - https://github.com/mohammeddanishrabbani/Legal-Simplification-mrabbani

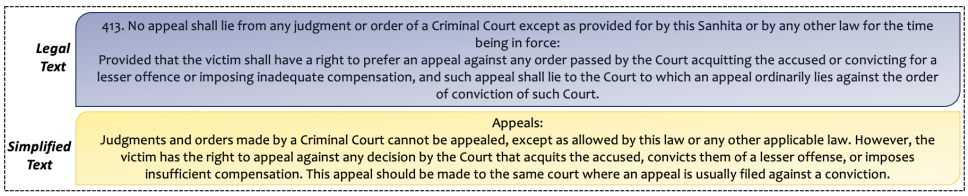

Figure 1: An instance from the *SIMPLE-LAW* dataset

These legal experts were given 50 random raw texts from the collected data and were asked to simplify these texts such that it can be understood by non-legal laypersons. Subsequently, the two legal experts were made to discuss with each other and asked to choose final 50 annotated samples with mutual agreement from a cohort of 100 samples which were ultimately included as part of the gold-standard dataset.

3.3. SIMPLE-LAW Dataset

The *SIMPLE-LAW* dataset comprises of 6,230 pairs of complex legal clauses and their corresponding simplified versions. Each data point consists of a source (i.e., legal_text) and a target (i.e., simplified_text). An instance from the dataset is provided in Figure 1. The *average word count* for the original legal texts in the dataset is approximately 292.72 words, whereas the corresponding simplified texts contain an average of 137.23 words. This indicates a substantial reduction in textual length. This compression ratio highlights the degree of abstraction and linguistic simplification applied during the transformation process. We surmise that *SIMPLE-LAW* sets a foundational benchmark for building accessible legal technologies, especially for users without legal expertise.

Quantitative Analysis. Given that the simplified texts in *SIMPLE-LAW* are synthetically generated, we took deliberate measures to ensure their high quality. To rigorously assess their readability and simplicity, we employed a suite of standard automated evaluation metrics, as detailed below.

• **The Flesch Reading Ease score (FRE)** (Flesch, 1948) evaluates how easy a text is to read based on sentence length and syllable complexity. Higher scores indicate simpler, more accessible language.

• **The Flesch-Kincaid Grade Level (FKGL)** (Kincaid et al., 1975) is an extension of FRE, adapted for use in educational and military setups. It is designed to indicate the U.S. school grade level required to understand a given piece of text. It is one of the most widely used readability metrics and is particularly useful when evaluating educational materials, legal texts, web content, or any writing where the target audience's reading proficiency matters.

• **Simple Measure of Gobbledygook (SMOG)** (McLaughlin, 1969) predicts the years of education required to comprehend a text, based specifically on polysyllabic words. SMOG is particularly effective for technical domains like law. A drop in SMOG score after simplification suggests reduced legalese and jargon.

• **Dale–Chall Readability Score** (Chall and Dale, 1995) is a classic readability measure designed to assess the difficulty level of a text based on familiar vocabulary and sentence length.

Table 1: Quantitative analysis of *SIMPLE-LAW* based on several readability & simplicity metrics; comparison with *BillSum* is also reported

Metric	Original Text	Simplified Text	BillSum
Flesch Reading Ease score (↑)	16.34	**32.30**	4.04
Flesch-Kincaid Grade Level (↓)	23.56	**15.08**	22.76
Simple Measure of Gobbledygook (↓)	16.96	**13.99**	15.47
Dale–Chall Readability Score (↓)	9.83	**9.64**	11.37

The raw legal texts exhibit extremely low readability as evidenced in Table 1 by a FRE score of 16.34, a FKGL of 23.56 and a SMOG Index of 16.97, indicating that comprehension of these texts require graduate-level literacy. The Dale–Chall score of 9.83 further confirms the prevalence of rare or domain-specific terminology. In contrast, the simplified texts achieve meaningful improvements across all metrics. The FRE improves to 32.31, moving toward more accessible reading levels, while the FKGL and SMOG scores decrease to 15.08 and 13.99, respectively, corresponding to undergraduate-level readability. Although the Dale–Chall score remains relatively high at 9.64, this slight reduction still indicates marginal simplification in vocabulary usage.

Additionally, we also perform evaluation to verify semantic preservation (Reimers and Gurevych, 2019) and entailment consistency between the original legal text and generated simplified text pairs in the dataset. (Ujwal et al., 2024) uses a sentence transformer cross-encoder model all-MiniLM-L6-v2 (Reimers, 2021) trained on NLI tasks (Williams et al., 2018) to ensure semantic similarity. We adopted this method to convert the legal texts and its corresponding simplified texts into vector embeddings and computed the cosine similarity to

Table 2: Quantitative analysis of *SIMPLE-LAW* based on several semantics & entailment metrics

Metric	Value
Average Cosine Similarity	0.8064
Semantic Similarity Count (%)	93.71%
Average Entailment Score	0.7438
Entailment Count (%)	80.45%
Joint Count (%)	75.70%

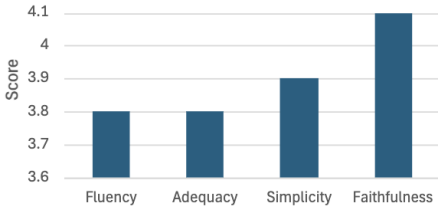

Figure 2: Qualitative analysis of *SIMPLE-LAW* dataset on human evaluation metrics

establish semantic similarity. The results are reported in Table 2 with a threshold of 0.65 (as used in (Ujwal et al., 2024)). The *Average Cosine Similarity* was observed to be 0.8064, suggesting a robust semantic alignment between the original and simplified texts. The high *Semantic Similarity Count* of 93.71% (proportion of pairs exceeding a 0.65 similarity threshold) further corroborates that the simplified texts preserve the informational essence of the legal inputs, even when surface-level lexical changes are introduced. For measuring entailment consistency, we utilized a NLI model, i.e., facebook/bart-large-mnli (Yin et al., 2019). Entailment captures whether the simplified version preserves factual truths present in the original text. We used the simplified text as hypothesis and raw legal text as premise to evaluate the probability of entailment vs. contradiction, discarding the "neutral" label (Lewis et al., 2019). The results are reported in Table 2. The *Average Entailment Score* computed as the mean probability that a simplified sentence entails its corresponding legal sentence stands at 0.7438. This relatively high value indicates that the simplifications produced are not only structurally modified but also largely consistent with the original legal meaning. Importantly, an entailment threshold of 0.5 was used to classify a simplification as logically consistent. Under this criterion, 80.45% of the dataset samples exhibit *entailment*, reflecting a strong preservation of core legal semantics. The *Joint Count* representing the proportion of samples satisfying both the entailment and semantic similarity thresholds is 75.70%. This compound metric is particularly significant as it reflects the dual fidelity of the simplified outputs both in terms of logical consistency and semantic preservation.

Qualitative Analysis. We conducted human evaluation to assess the quality of the simplified legal texts along key linguistic and semantic dimensions. 100 random samples were chosen from the *SIMPLE-LAW* dataset and were presented to two legal experts and one layperson from the author's affiliation. They were asked to rate (1 (poor) - 5 (excellent)) these samples based on four criteria: **(i)** *Fluency* measures grammatical correctness and naturalness of the output; **(ii)** *Adequacy* assesses whether all essential information from the original legal text is preserved; **(iii)** *Simplicity* evaluates the ease of comprehension for a layperson; and **(iv)** *Faithfulness* determines whether the legal meaning remains intact without introducing hallucinations or distortions. The results are reported in Figure 2. The relatively high fluency score indicates that the outputs are generally well-formed, grammatically correct, and coherent. Adequacy metric scored modestly revealing a trade-off between simplification and content preservation, which is particularly critical in legal domains where nuance and specificity are paramount. The model performs strongly on simplicity with annotators noting improved accessibility and reduced legalese. The highest score was achieved in faithfulness which measures the factual consistency between the original and simplified text. A score above 4.0 confirms that there is low hallucination in the simplified text.

Comparative Analysis. We extensively looked for existing datasets in legal and other domains for our specific use-case. BillSum (Kornilova and Eidelman, 2019) contains US Congressional bills along with human-written summaries. These summaries could be adapted for simplification since summaries are often written in accessible language. LEDGER (Tuggener et al., 2020) contains over 60 categories of clause-level segments from commercial contracts. This dataset is primarily used for classification and cannot be used for our task. We found various general domain datasets for simplification such as WikiLarge (Zhang and Lapata, 2017), ASSET (Alva-Manchego et al., 2020b) which are good for base simplification model pre-training but might fail to capture the complexity that exists in legal text. Newsela (Xu et al., 2015) contains news articles rewritten by professionals at multiple readability levels. But, it has small number of samples. We compared *SIMPLE-LAW* with BillSum dataset on standard readability metrics. The results summarized in Table 1, reveals that *SIMPLE-LAW* consistently outperforms BillSum across all metrics. Notably, *SIMPLE-LAW* achieves a FRE score of 32.3, significantly higher than BillSum's 4.042, indicating a more accessible reading experience. Furthermore, the lower FKGL and SMOG index suggest that texts in *SIMPLE-LAW* are suitable for readers with fewer years of formal education. The Dale–Chall score also reflects greater lexical simplicity in *SIMPLE-LAW* (9.64 vs. 11.37).

Table 3: Comparison of existing text simplification dataset

Dataset	Domain	#Samples	Parallel	Simplification Type
BillSum (Kornilova and Eidelman, 2019)	US Legislation	∼23,000	Yes	Bill → Summary (Layperson)
WikiLarge (Zhang and Lapata, 2017)	Wikipedia	∼296,000	Yes	Complex → Simple Sentences
ASSET (Alva-Manchego et al., 2020b)	Wikipedia	2,359	Yes (multi-ref)	Sentence-level Simplification
LEDGAR (Tuggener et al., 2020)	Contract Clauses	∼850,000	No	Classification / Clause Reuse
Newsela (Xu et al., 2015)	News Articles	1,191	Yes	Multiple Simplification Levels
SIMPLE-LAW (Ours)	Multi-Jurisdiction Legal Data	6,230	Yes	Legal → Layperson

4. Methodology

This section delineates the experimental framework and implementation details for evaluating selected LLMs to compare the performance of fine-tuned and in-context variants on the legal simplification, with a focus on improving layperson readability.

4.1. Benchmark Setup

The target task is legal text simplification aimed at making statutory and legal language more comprehensible to a general audience. The benchmark assumes a supervised sequence-to-sequence paradigm where input is a complex legal text and output is its simplified version. To explore the trade-off between efficiency and performance, we selected models ranging from 1B to 7B parameters, all adapted for instruction-following tasks and deployed via the Unsloth framework (Han et al., 2023). The models include:

(1) `Gemma-3-4bit` (Unsloth, 2024) (LoRA $r = 8$) - Chosen for its compact size and instruction-tuned design. Suitable for multilingual legal corpora like Indian and UK laws. Gemma's 4B footprint offers a strong balance between performance and deployability.

(2) `Llama-3.2-1B-Instruct-bnb-4bit` (Unsloth, 2025a) (LoRA $r = 16$) - The smallest model selected to explore the minimum viable capacity for legal simplification under extreme parameter constraints.

(3) `Llama-3.2-3B` (Unsloth, 2025b) (LoRA $r = 16$) - A mid-range LLaMA model known for strong generalization. It has been selected to verify the necessity of instruction-tuned models.

(4) `Mistral-7b-instruct-v0.3-bnb-4bit` (Unsloth, 2025c) (LoRA $r = 8$) - Selected for its strong instruction-tuning and efficient decoder design and serves as a high-performing baseline within the 7B.

(5) `Qwen2.5-1.5B-Instruct-bnb-4bit` (Unsloth, 2025d) (LoRA $r = 16$) - A mid-size multilingual model, well-suited for diverse legal inputs. It allows us to test whether 1.5B parameters are sufficient when paired with high-rank LoRA adapters.

(6) `Qwen2.5-7B` (Unsloth, 2025e) with $r = 16$ - Included to study the impact of LoRA rank (r=16) on model capacity and legal language understanding at the 7B scale.

All models were quantized to 4-bit precision and fine-tuned using the QLoRA technique (Dettmers et al., 2023), which enables efficient training with low memory overhead and were evaluated on In-Context Learning (ICL) and Fine-Tuning (FT) settings.

Fine-tuning Strategy. We utilized QLoRA (Dettmers et al., 2023) which enables low-rank adaptation (Hu et al., 2021) (LoRA) of 4-bit quantized models, significantly reducing the computational overhead while maintaining performance. Each model is fine-tuned on the entire training dataset in a supervised manner, learning to map complex legal input texts to their corresponding simplified outputs. LoRA ranks are configured (e.g., $r = 8$ or $r = 16$) depending on the model capacity, allowing effective adaptation with minimal trainable parameters.

In-context Learning Strategy. In contrast, the ICL strategy leverages pretrained models without updating their weights. We have tried a range of k-shot prompting techniques ranging from 0-2. This setup mimics human-like learning and has gained traction as an effective alternative for downstream tasks (Brown et al., 2020; Chen and Wang, 2022). While ICL offers deployment flexibility without requiring model retraining, it is often sensitive to prompt structure and model size.

4.2. Implementation Details

80% of the dataset was used as training set for FT/ICL using the Unsloth (Han et al., 2023) library. We adopted the **Unsloth** implementation due to its practical advantages (Saputra, 2025) for scalable and efficient experimentation with instruction-tuned LLMs across various models (1B–7B) on a single consumer-grade NVIDIA RTX 2080 Ti GPU, without compromising training stability or output quality. FT was performed with a batch size of 4, gradient accumulation steps of 1, `AdamW_8bit` (Dettmers et al., 2022) optimizer, learning rate of 2e-4, loRA alpha equals rank, and with 5 warmup steps. LoRA ranks were set to 8 or 16 as specified. Inputs were tokenized using model-specific tokenizers with a maximum length of 8192 tokens. During inference, we employed a controlled decoding strategy to ensure fluency, diversity, and factual consistency in the output with a maximum generation length of 8192 tokens to accommodate longer legal inputs and outputs. A moderate temperature of 0.7 was

used to balance creativity and determinism, while top-k sampling (with $k = 50$) and top-p sampling (with $p = 0.9$) were applied to restrict the sampling space to high-probability tokens, promoting both diversity and coherence. To discourage verbose or repetitive completions, we introduced a repetition penalty of 1.2 and enforced a no-repeat n-gram size of 3, ensuring that trigrams do not repeat within the output. The generation process was terminated upon encountering the model's end-of-sequence token (`eos_token_id`). This configuration was consistent across all models and experiments to ensure comparability of results.

4.3. Evaluation Metrics

In this section, we outline several evaluation strategies and metrics to assess the quality of the generated output by different models.

Automated Metrics. We have evaluated the models for FT and ICL on various standard automated metrics for text generation such as *BERTScore* (Zhang et al., 2020) and *ROUGE* metrics (Lin, 2004). In addition to standard automated evaluation metrics, we employ *SARI* (System output Against References and against the Input) (Xu et al., 2016), a widely recognized metric specifically designed for text simplification tasks. Unlike conventional metrics that rely solely on comparing system outputs with reference simplifications, SARI evaluates the quality of simplification by explicitly rewarding the system's ability to perform appropriate additions, deletions, and retentions. It does so by simultaneously comparing the system output with both the original complex input and multiple reference simplifications. This allows SARI to effectively capture the trade-off between content preservation and readability enhancement, making it particularly suitable for simplification tasks.

Hallucination Analysis. Additionally, we also performed evaluation to analyze how each of these models hallucinated while generating simplified versions of the legal text, i.e., measuring the *factuality* of the generated simplified texts. We utilized a sentence transformer cross-encoder model[2] similar to (Ujwal et al., 2024) which outputs a probability between 0 and 1 where 0 indicates hallucination and 1 being factually consistent. We used the hypothesis as the ground truth and the premise as the generated simplified text by the model.

Human Evaluation Metrics. We conducted human evaluation to assess the quality of the simplified legal text generated from the fine-tuned models since the fine-tuned models performed better when judged on automated metrics. 50 random samples were chosen from the *SIMPLE-LAW* dataset along with responses from all the FT models and were presented to one legal expert and two non-legal layperson from the author's affiliation. They were asked to rate (1 (poor) - 5 (excellent)) these samples based on four criteria mentioned in Section 3.3.

[2] https://huggingface.co/vectara/hallucination_evaluation_model

5. Results and Analysis

This section presents a comprehensive analysis of the experimental results obtained by evaluating multiple QLoRA-LLMs across ICT and FT settings for the legal text simplification task.

5.1. Experimental Results

The evaluation of the models focuses on semantic preservation, syntactic fidelity, factual accuracy, and simplification efficacy measured using BERTScore, ROUGE, SARI, and a custom hallucination-based factuality metric discussed above. The results have been logged in Table 4. As observed, the *Mistral-7B* model demonstrated overall superior performance with the highest BERTScore-F1 at 0.9289 in the fine-tuned setting. Also, it achieved top scores in all the ROUGE-1/2/L/Lsum metrics suggesting an excellent lexical and syntactic overlap with the reference simplifications. In terms of SARI, it achieved a notable 66.29, outperforming other models by a significant margin, indicating high-quality simplifications in terms of appropriate additions, deletions, and keeps. On the other hand, *LLaMA-3.2-1B* variant showed an interesting trend with the fine-tuned variant achieving strong results (BERTScore-F1 = 0.9031, SARI = 50.06, ROUGE-1 = 0.5888) despite being significantly smaller. The in-context learning variants, especially 1-shot and 2-shot performed competitively in semantic similarity (BERTScore-F1 = 0.85), but fell short in SARI with score less than 43 indicating weaker structural simplification. It also outperformed *LLaMA-3.2-3B* model thereby showing that instruction tuned models perform better. *Gemma-3-4B* model yielded strong results with in-context learning strategy. ICL-2-shot showed effective balance approaching fine-tuned Mistral's performance. The fine-tuned variant in this case did not substantially outperform in-context ones, possibly indicating Gemma's limited gains from low-resource fine-tuning compared to more instruction-aligned models. Both the models from the *Qwen* family did not perform well on legal text simplification task. This possibly can be attributed to pretraining on code and mathematical reasoning datasets which limits their alignment with legal narrative structures. As observed throughout, across all models, fine-tuning with QLoRA yields substantial improvements over in-context learning. Mistral-7B shows an increase

Table 4: Quantitative analysis of different models on the *SIMPLE-LAW* dataset. The best performing strategy for each model is highlighted in blue and the second best in red

Model(s)	Strategy	BERTScore			ROUGE Score				SARI	Factual Score
		p	r	f	1	2	L	Lsum		
LLama-3.2-1B	ICL-0-shot	0.83	0.87	0.85	0.32	0.16	0.23	0.29	42.65	0.55
	ICL-1-shot	0.84	0.86	0.85	0.33	0.16	0.23	0.29	41.54	0.58
	ICL-2-shot	0.83	0.87	0.85	0.27	0.13	0.20	0.24	42.32	0.73
	FT	0.90	0.90	0.90	0.59	0.34	0.47	0.53	50.06	0.60
Qwen-2.5-1.5B	ICL-0-shot	0.86	0.85	0.86	0.29	0.09	0.20	0.26	35.39	0.20
	ICL-1-shot	0.84	0.85	0.85	0.29	0.10	0.20	0.26	36.40	0.30
	ICL-2-shot	0.85	0.85	0.85	0.26	0.08	0.19	0.23	35.07	0.26
	FT	0.88	0.88	0.88	0.42	0.19	0.32	0.37	42.70	0.43
LLama-3.2-3B	ICL-0-shot	0.81	0.82	0.81	0.13	0.02	0.09	0.12	29.56	0.09
	ICL-1-shot	0.81	0.83	0.82	0.17	0.04	0.11	0.15	31.39	0.11
	ICL-2-shot	0.81	0.83	0.82	0.18	0.03	0.11	0.16	31.15	0.13
	FT	0.88	0.89	0.88	0.43	0.23	0.35	0.40	41.77	0.35
Gemma-3-4B	ICL-0-shot	0.82	0.88	0.85	0.39	0.15	0.24	0.35	42.97	0.28
	ICL-1-shot	0.84	0.89	0.87	0.46	0.21	0.32	0.41	45.00	0.38
	ICL-2-shot	0.88	0.90	0.89	0.51	0.24	0.37	0.46	45.00	0.45
	FT	0.82	0.88	0.85	0.38	0.15	0.24	0.34	42.44	0.27
Qwen-2.5-7B	ICL-0-shot	0.82	0.84	0.83	0.17	0.03	0.11	0.15	31.28	0.06
	ICL-1-shot	0.78	0.82	0.80	0.16	0.03	0.10	0.14	30.52	0.14
	ICL-2-shot	0.77	0.81	0.79	0.11	0.02	0.07	0.10	29.61	0.06
	FT	0.82	0.87	0.85	0.31	0.13	0.23	0.28	37.97	0.29
Mistral-7B	ICL-0-shot	0.89	0.91	0.90	0.56	0.32	0.44	0.51	51.23	0.60
	ICL-1-shot	0.90	0.92	0.91	0.60	0.37	0.50	0.56	53.28	0.58
	ICL-2-shot	0.90	0.92	0.91	0.59	0.36	0.50	0.55	53.32	0.57
	FT	0.93	0.93	0.93	0.73	0.57	0.67	0.70	66.29	0.66

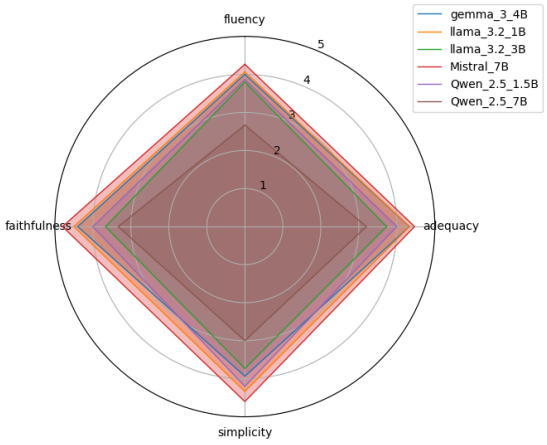

Figure 3: Human evaluation scores of different models in fine-tuning setting

in SARI from 53.3 (in-context-1–2-shot) to 66.29 (fine-tune). LLaMA-3.2-1B improves its ROUGE-1 score from 0.32 (in-context) to 0.59 (fine-tune) and its SARI from 42.6 to 50.06. Qwen-2.5-1.5B and Gemma-3-4B also show notable SARI gains post-finetuning.

Hallucination Analysis. The factuality scores derived across both ICL and FT strategies reveal notable trends in hallucination behavior. Amongst the evaluated models, *Mistral-7B* demonstrated consistently high factual accuracy across all prompting strategies, with the FT variant achieving the highest factuality score of 0.664, indicating its robustness against hallucination. Interestingly, *LLama-3.2-1B (instruction-tuned)* outperformed its larger 3B (not-instruction tuned) counterpart in ICL settings, particularly with 2-shot prompting (0.733) suggesting that instruction-tuned smaller model may generalize better with few-shot examples when appropriately tuned. This can be confirmed with *Qwen-2.5-7B (not instruction-tuned)* which exhibited the weakest performance, especially in ICL modes, where scores dropped as low as 0.061, highlighting a tendency toward hallucinations under few-shot prompting. However, fine-tuning improved its performance to 0.291, although still trailing behind other models. However, *Qwen-2.5-1.5B (instruction-tuned)* performed better which showed moderate gains through fine-tuning (0.433), outperforming its ICL configurations. Collectively, these observations suggest that model-specific architectural differences, parameter scales, and the choice of strategy (ICL vs FT) substantially impact factual consistency, with fine-tuning generally yielding better factual adherence except in specific cases like *LLama-1B* and *Gemma-4B*, where 2-shot prompting proved superior.

Human Evaluation. Based on the human evaluation scores (shown in Figure 3) several trends emerge that compare performance of the fine-tuned models on the legal text simplification task. *Mistral-7B* consistently outperformed other models across all metrics, particularly excelling in faithfulness (4.8) and simplicity (4.6), indicating that it not only preserves the meaning of the original legal text but also expresses it in a more comprehensible manner. *LLaMA-3.2-1B* followed closely, with strong performance in simplicity (4.33) and faithfulness (4.5) suggesting that smaller-scale models can also be competitive when well-optimized.

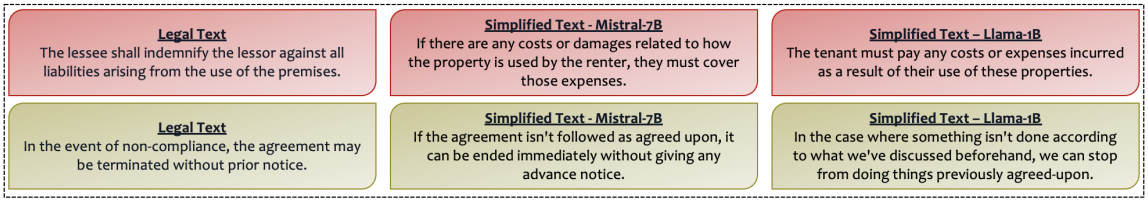

Figure 4: Sample output generated by different models in fine-tuning setting

In contrast, *LLaMA-3.2-3B* and *Qwen-2.5-7B* underperformed, particularly the latter, which scored the lowest across all dimensions, especially in fluency & simplicity possibly indicating the importance of instruction-tuning of models as *Qwen-2.5-7B* and *LLaMA-3.2-3B* are not models of *instruct* category. Interestingly, *Gemma-3-4B* and *Qwen-2.5-1.5B* achieved moderate scores across metrics, with Gemma slightly edging out in adequacy, reflecting a balanced but not outstanding capability. These results underscore the significance of not just model size, but also the architecture and instruction-tuning quality in determining simplification performance. Moreover, they highlight that larger instruction-tuned models like *Mistral-7B* can better grasp and translate complex legal language into simpler forms, aligning closely with human expectations of clarity and accuracy.

Qualitative Analysis. Figure 4 presents a qualitative comparison between outputs generated by two best fine-tuned models: *Mistral-7B* and *LLaMA-1B*. In the first example, *Mistral-7B* produces a more fluent and idiomatic text which is natural, coherent, and closely aligned with layperson comprehension, using familiar terms such as "*costs or damages*" and "*renter*". In contrast, *LLaMA-1B*'s output while correct and structurally sound, remains closer to the original legal register and includes slightly more formal/legal terminology such as "*incurred*" and "*these properties*". This suggests that *Mistral-7B* may generalize better. In the second example, *Mistral-7B* output is accurate, fluent, and preserves the immediacy of termination implied in the original clause. *LLaMA-1B*'s output, however, exhibits signs of grammatical awkwardness introducing informal phrasing ("*we've discussed beforehand*") and includes an unclear construction ("*stop from doing things*"), which may reduce the understanding. Overall, the qualitative analysis highlights that the *Mistral-7B* (FT) consistently delivers more accurate, grammatically coherent, and user-friendly simplifications than *LLaMA-1B*.

5.2. Findings to Research Questions

In this section, we provide answers to our research questions backed by empirical findings.

RQ(i): Can industry-grade LLMs be used to curate domain-specific datasets? Yes, industry-grade LLMs (e.g., GPT-3.5-Turbo) demonstrated strong utility in curating high-quality domain-specific datasets, especially when guided through structured prompting strategies such as 1-shot, 2-shot, and chain-of-thought (CoT) (Wei et al., 2022). As evident, *SIMPLE-LAW* dataset was generated using *GPT-3.5-Turbo* through templated prompts tailored to simplify legal clauses and the manual and automated evaluation of the generated simplifications confirmed linguistic fluency, syntactic correctness, and general adherence to factual consistency (seen in Table 1 & 3). The overall dataset quality was robust enough to support different model fine-tuning. The performance of the fine-tuned models such as *Mistral-7B* further validates the dataset's utility with a high SARI, ROUGE, and BERTScore values suggesting that the curated simplifications offered a strong learning signal. Thus, industry-grade LLMs can serve as efficient, scalable tools for bootstrapping task-specific datasets in low-resource domains such as legal text simplification.

RQ(ii): How effective is the use of semi-synthetic datasets in mimicking the real-world legal simplification task? Semi-synthetic datasets, when generated via high-quality prompting strategies, can effectively replicate the complexity and nuance of real-world legal simplification tasks, albeit with some limitations regarding factual consistency in low-capacity models. Fine-tuned models trained exclusively on the semi-synthetic dataset exhibited substantial gains across all evaluation metrics compared to in-context variants, indicating effective learning from synthetic supervision.

RQ(iii): How do fine-tuned LLMs perform compared to the in-context learning strategy for the legal simplification task? Fine-tuned LLMs, particularly those trained using QLoRA, significantly outperformed in-context learning strategies across all measured dimensions. *Mistral-7B* showed a stark improvement in simplification quality when fine-tuned compared to its best ICL variant. *LLaMA-3.2-1B*, despite its small size, demonstrated strong gains. Parameter-efficient techniques such as QLoRA allows scalable fine-tuning even for resource-constrained environments while outperforming prompt-based alternatives.

6. Conclusion and Future Works

In this work, we introduced *SIMPLE-LAW*, a novel dataset comprising of 6,230 high-quality pairs of complex legal text and their corresponding simplified versions, generated with a LLM using multiple in-context learning strategies. The dataset is specifically curated to support research in legal text simplification, a task which lacked a standard dataset. The dataset was further benchmarked with a suite of Unsloth LLMs across multiple model sizes, using rigorous evaluation metrics both in fine-tuning and in-context learning strategies. Our findings indicate that lightweight models achieve competitive simplification quality with reduced inference costs, suggesting practical viability for deployment in legal aid systems.

Future works will include (i) enhancing factual consistency and legal soundness through supervised alignment and reinforcement learning with human feedback (RLHF) (Lambert, 2025); (ii) expanding the dataset to encompass multilingual legal corpora and diverse legal subdomains can broaden the model's generalization etc. We surmise that this work furthers the initiative to bring the research community closer to trustworthy, accessible, and scalable AI systems for the legal domain.

7. Bibliographical References

Mistral AI. 2023. Mistral 7b. https://mistral.ai/news/announcing-mistral-7b. Accessed: 2025-05-19.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2020b. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679.

Mark Athugodage, Olga Mitrofanove, and Vadim Gudkov. 2024. Transfer learning for Russian legal text simplification. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 59–69, Torino, Italia. ELRA and ICCL.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Mert Cemri, Tolga Çukur, and Aykut Koç. 2022. Unsupervised simplification of legal texts.

Jeanne S. Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge, MA.

Yinbo Chen and Xiaolong Wang. 2022. Transformers as meta-learners for implicit neural representations.

Google DeepMind. 2024. Gemma: Open models based on gemini research and technology. https://ai.google.dev/gemma. Accessed: 2025-05-19.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Anand Devaraj and et al. 2023. Medsimplify: A dataset for simplification of medical texts. In *Proceedings of ACL BioNLP*.

European Commission. n.d. Clear writing in the european commission. Accessed: 2025-05-18.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Isabel Gallegos. 2022. The right to remain plain: Summarization and simplification of legal documents.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nandakishore Kambhatla. 2022. Text simplification for legal domain: Insights and challenges. In

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 296–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

John Gibbons. 2003. *Forensic Linguistics: An Introduction to Language in the Justice System*. Blackwell Publishing, Oxford, UK.

Aaron Grattafiori and et al. 2024. Llama 3 herd of models. https://arxiv.org/abs/2407.21783. Accessed: 2025-05-19.

D. Han, M. Han, and the Unsloth Team. 2023. Unsloth. https://github.com/unslothai/unsloth.

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.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report Research Branch Report 8-75, Naval Air Station Memphis, Chief of Naval Technical Training.

Anastassia Kornilova and Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Nathan Lambert. 2025. Reinforcement learning from human feedback.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Lighteval. 2023. legal_summarization.

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.

Aman Madaan and et al. 2022. Text simplification with human-in-the-loop feedback. In *ACL*.

Louis Martin, Benoît Sagot, Éric de la Clergerie, and Antoine Bordes. 2020. Controllable sentence simplification.

G. Harry McLaughlin. 1969. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

Junya Nishihara and et al. 2022. Readability-controllable text simplification with enhanced reward mechanism. In *EMNLP*.

OpenAI. 2023. Gpt-3.5-turbo technical report. https://platform.openai.com/docs/models/gpt-3.5-turbo. Accessed: 2025-05-23.

PWOnlyIAS. 2025. Simplifying legal language. https://pwonlyias.com/editorial-analysis/simplifying-legal-language/. Accessed: 2025-05-09.

Nils Reimers. 2021. all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2025-05-23.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Aditya G. Saputra. 2025. Medical diagnosis chatbot assistant. https://medium.com/@adityagofi/medical-diagnosis-chatbot-assistant-71fb0d33dc

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298.

Qwen Team. 2024. Qwen2.5: A party of foundation models! https://qwenlm.github.io/blog/qwen2.5/. Accessed: 2025-05-19.

Don Tuggener, Pius von Däniken, Mark Cieliebak, and Hans Hofmann. 2020. Ledgar: A large-scale multi-label dataset for document classification in the legal domain. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1235–1241.

Utkarsh Ujwal, Sai Sri Harsha Surampudi, Sayantan Mitra, and Tulika Saha. 2024. "reasoning before responding": Towards legal long-form question answering with interpretability. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4922–4930, New York, NY, USA. Association for Computing Machinery.

United Nations Development Programme. 2022. Access to justice. https://www.undp.org/justice/access-to-justice. Accessed: 2025-05-18.

Unsloth. 2024. unsloth/gemma-3-4b-it. https://huggingface.co/unsloth/gemma-3-4b-it. Accessed: 2025-05-15.

Unsloth. 2025a. unsloth/llama-3.2-1b-instruct-bnb-4bit. https://huggingface.co/unsloth/Llama-3.2-1B-Instruct-bnb-4bit. Accessed: 2025-05-16.

Unsloth. 2025b. unsloth/llama-3.2-3b. https://huggingface.co/unsloth/Llama-3.2-3B. Accessed: 2025-05-16.

Unsloth. 2025c. unsloth/mistral-7b-instruct-v0.3-bnb-4bit. https://huggingface.co/unsloth/mistral-7b-instruct-v0.3-bnb-4bit. Accessed: 2025-05-16.

Unsloth. 2025d. unsloth/qwen2.5-1.5b-instruct-bnb-4bit. https://huggingface.co/unsloth/Qwen2.5-1.5B-Instruct-bnb-4bit. Accessed: 2025-05-16.

Unsloth. 2025e. unsloth/qwen2.5-7b. https://huggingface.co/unsloth/Qwen2.5-7B. Accessed: 2025-05-16.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. In *Proceedings of NAACL*.

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.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

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

Xinyu Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.