SINBRIEF: A Hybrid Framework for Abstractive Text Summarisation of Sinhala Legal Documents

Minduli Lasandi[♠] and **Nevidu Jayatilleke**[♣]
[♠]School of Computing, Informatics Institute of Technology, Sri Lanka
[♣]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
minduli.20220374@iit.ac.lk, nevidu.25@cse.mrt.ac.lk

Abstract

Legal document summarisation in low-resource languages presents significant challenges due to the scarcity of annotated data and the complexity of domain-specific terminology. This paper presents SINBRIEF, a hybrid abstractive summarisation framework for Sinhala legal documents that does not require human-annotated training data. The proposed framework combines domain-aware word graph construction with neural sentence scoring to generate abstractive summaries from Sinhala legal text. Five sentence scoring models are evaluated within the framework: `mBert`, `Llama 3.1`, `Falcon 7B`, `Laser`, and a continually pretrained `Llama` model domain-adapted to Sinhala legal text. The framework is evaluated on a Sinhala legal corpus using reference-free metrics, including Coverage, Density, Compression Ratio, SummaC, and Self-BertScore. Experimental results demonstrate that SINBRIEF produces summaries with lower lexical overlap than extractive baselines while maintaining factual consistency, demonstrating the viability of hybrid, largely annotation-free abstractive summarisation for low-resource legal NLP tasks.

Keywords: Low-resource NLP, Text summarisation, Legal documents

1 Introduction

Legal documents are often lengthy, complex, and linguistically dense, making them difficult for the general public to interpret efficiently (Jayatilleke and Weerasinghe, 2025). In Sri Lanka, where legal records are predominantly written in Sinhala, the absence of automated summarisation tools further limits accessibility and information retrieval (Jayawardane, 2022). As a result, legal practitioners, researchers, and citizens must manually navigate extensive textual content to extract relevant information, which can be time-consuming and cognitively demanding.

Automatic abstractive text summarisation offers a promising solution by generating concise, coherent summaries that capture the semantic essence by producing novel sentences that capture the semantic meaning of a source document (Jayatilleke et al., 2024). Unlike extractive summarisation methods, which select sentences directly from the original text (Azam et al., 2025), abstractive approaches enable deeper semantic compression and improved readability. These properties are particularly beneficial in legal contexts, where extracting sentences may preserve redundancy and structural complexity rather than improving clarity.

While significant progress has been achieved in abstractive summarisation for high-resource languages such as English, limited attention has been given to low-resource languages due to the lack of resources (Munaf et al., 2024). The scarcity of large-scale annotated datasets and domain-specific corpora poses substantial challenges for supervised learning approaches, particularly in specialised domains such as legal documentation (Ariai et al., 2025). In low-resource settings, methods that minimise reliance on human annotation offer a promising alternative, as they eliminate the dependency on parallel document–summary datasets while still enabling semantic content compression (Nikolov, 2020). However, research on abstractive summarisation for Sinhala legal texts remains largely unexplored.

To address this gap, this research proposes SINBRIEF, a hybrid abstractive text summarisation framework for Sinhala legal documents. The framework is largely annotation-free, with the core summarisation pipeline operating without supervision and the scoring component employing weak supervision through constraint-guided label construction, requiring no human annotation. The proposed approach incorporates domain-aware word graph construction and sentence scoring to generate

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 252–265
November 6-7, 2026 ©2026 Association for Computational Linguistics

coherent and informative summaries.

2 Existing Work

Text summarisation has been widely studied in Natural Language Processing, with research spanning extractive, abstractive, supervised, and unsupervised methodologies. These efforts have driven significant methodological advancements across multiple languages and domains.

2.1 BenSumm

Bensumm is described as the first unsupervised abstractive summarisation framework for Bengali documents (Chowdhury et al., 2021). This model adopts a graph-based, single-document approach that relies on POS tagging and a pre-trained language model named ULMFiT (Howard and Ruder, 2018) embeddings to perform sentence clustering, word-graph construction and sentence fusion. By generating and ranking candidate paths using a ranking strategy based on the work of Boudin and Morin (2013) within cluster-specific word graphs, the system produces abstractive summaries without requiring parallel training data.

In addition to the model, the authors released a human-annotated benchmark dataset consisting of 139 document–summary pairs, characterised by a high degree of abstractiveness. Experimental results demonstrated that BenSumm outperformed established unsupervised extractive baselines such as LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004) in terms of ROUGE (Lin, 2004) scores and human evaluation metrics.

SINBRIEF adopts BenSumm's general pipeline of sentence clustering, word graph construction, and graph-based candidate generation, but introduces three key novelties: (i) domain-aware edge weighting that assigns higher weights to edges involving Sinhala legal keywords; (ii) a weakly supervised neural sentence scoring model that replaces BenSumm's fixed ranking strategy with a trained binary classifier for legal sentence quality; and (iii) continual pre-training of Llama 3.1 8B on the Sinhala legal corpus to produce a domain-adapted backbone.

2.2 Summarisation of Legal Texts

The summarisation of legal texts has become increasingly important due to their complexity and length, and researchers have developed a variety of extractive and abstractive methods leveraging advanced NLP techniques and large language models to address these challenges.

M et al. (2024) explored extractive and abstractive summarisation of legal documents using transformer models and large language models such as GPT-4 (Achiam et al., 2023) and Llama-2 (Touvron et al., 2023), supporting multilingual output and applications like chatbots. Grover et al. (2003a) applied summarisation to legal judgments by classifying sentences according to their rhetorical roles, using NLP tools and linguistic features to guide summary extraction, while Chheda et al. (2025) also introduced a rhetorical role-based Extract-Explain-Abstract framework for generating abstractive summaries in low-resource legal settings.

Hachey and Grover (2005) explored machine learning approaches, including Naïve Bayes and maximum entropy, for extractive summarisation of legal texts, leveraging rhetorical status information to score sentences and structure summaries.

Jayatilleke and Weerasinghe (2025) proposed a hybrid approach for patent summarisation, combining LexRank for extractive sentence selection with a Bart model (Lewis et al., 2020) fine-tuned via LoRA (Hu et al., 2022) for abstractive summary generation and domain generalisation across patent fields. Other research includes the *PRODIGIT* system by Pont et al. (2023), which uses GPT-4 to generate issue-based abstractive summaries of Italian tax law rulings, the fuzzy logic-based headnote generation approach by Megala et al. (2014), the fine-tuning of Bart and Pegasus (Zhang et al., 2020) for legal summarisation by Kasar et al. (2025) and the summarisation of legal judgments by Grover et al. (2003b) through classifying sentences based on their argumentative roles using linguistic features.

2.3 Unsupervised Summarisation Approaches

Unsupervised summarisation methods aim to automatically generate concise, coherent summaries without requiring parallel document-summary datasets. By leveraging techniques ranging from graph-based representations to neural architectures and hybrid frameworks, these approaches enable effective summarisation in domains where annotated data is rare.

Nayeem et al. (2018) proposed an unsupervised abstractive summarisation model based on paraphrastic sentence fusion. The approach combines

word-graph generation with semantic ranking to produce more informative summaries. Liu et al. (2015) proposed an AMR-based abstractive summarisation framework that generates summaries by transforming semantic graphs, while Dohare et al. (2018) introduced an unsupervised approach that improves document-level graph construction and subgraph extraction for summary generation. Ramirez-Orta and Milios (2021) introduced an unsupervised extractive method that ranks sentences using graph centrality over pre-trained sentence embeddings to capture semantic similarity. These approaches demonstrate the progression from simple word-graph and AMR-based methods toward more sophisticated document-level graph construction and embedding-based ranking.

Chu and Liu (2019) proposed MeanSum, an end-to-end neural model for unsupervised abstractive summarisation that generates summaries by decoding the mean representation of input documents. Zhou and Rush (2019) proposed an unsupervised summarisation approach based on contextual matching using pre-trained language models, while Coavoux et al. (2019) introduced an aspect-aware neural model that generates summaries by clustering document representations into semantic facets.

Other research includes Wu et al. (2022), which proposed an unsupervised abstractive summarisation framework to identify salient content using frequency-based semantic units, Tang et al. (2023), which applies a topic-guided co-attention transformer for extreme multimodal summarisation, Song et al. (2022), an unsupervised abstractive model leveraging Wasserstein distance for opinion summarisation, and Zhao et al. (2020), a graph-based method that combines sentence representations and spectral clustering for multi-document summarisation. These methods show that effective abstractive summarisation can be achieved without massive pre-trained models, making them suitable for low-resource domains.

3 Methodology

The proposed framework follows an hybrid abstractive summarisation approach designed for Sinhala legal texts. The method begins with sentence extraction and clustering, followed by graph-based representation to capture structural and semantic relationships within the document. Abstractive candidate sentences are then generated using a word graph-based fusion mechanism. To ensure the quality of the generated summaries, a scoring component is employed to rank candidate sentences. The top-ranked candidates are finally selected and combined to produce the final summary.

3.1 Data

This study utilises the `SinhaLegal` dataset, comprising 1206 legal documents of Acts and Bills (Lasandi and Jayatilleke, 2026). The dataset was partitioned into 80% training and 20% testing using a fixed random seed (42), resulting in 964 training documents and 242 testing documents.

3.2 Sentence Extraction and Clustering

The input document is segmented into sentences by splitting on punctuation marks, specifically periods, exclamation marks, question marks, and semicolons. Sentences shorter than 20 characters are discarded to eliminate incomplete fragments.

The retained sentences are then encoded into dense vector representations for semantic clustering. The choice of encoder varies depending on the scoring model used in the pipeline. For `mBert`-based and the continually pre-trained `Llama` variants, sentence representations are obtained from transformer models, while `Laser` employs its dedicated multilingual *encode_sentences()* API directly. For `mBert` (Devlin et al., 2019), the [CLS] token embedding from the final hidden layer serves as the sentence representation.

For `Llama 3.1 8B` (Grattafiori et al., 2024) and `Falcon 7B` (Almazrouei et al., 2023), which are causal decoder-only models with no [CLS] token, the hidden state of the last non-padding token from the final layer is used instead. Both large models are quantised to 4-bit NF4 precision via BitsAndBytes to enable inference within GPU memory constraints. For `Laser` (Artetxe and Schwenk, 2019), fixed-size 1024-dimensional embeddings are obtained directly from the frozen multilingual encoder without any token-level extraction.

Regardless of the encoder used, sentence representations are grouped into semantically coherent clusters using agglomerative clustering with average linkage and cosine similarity as the distance metric. The number of clusters is set to $k = \min(3, \lfloor N/5 \rfloor)$, where N is the total number of extracted sentences, capping the cluster count at 3, while reducing it for shorter documents with fewer than 15 sentences.

3.3 Word Graph Construction

For each sentence cluster, a directed word graph is constructed to capture co-occurrence relationships between words. Sentences within a cluster are first tokenised and POS-tagged using the *Sinling* library, which provides Sinhala-specific tokenisation and morphological analysis. Only tokens belonging to content-bearing POS categories (nouns, verbs, auxiliary verbs, proper nouns, and adjectives) are retained for graph construction. Each node in the graph represents a (word, POS tag) tuple, with directed edges connecting consecutive word pairs within a sentence. A START and END node are included to delimit valid sentence boundaries during candidate generation.

Edge weights are assigned using a domain-aware weighting scheme tailored to the legal domain. An edge receives a weight of 2.0 if either of its constituent nodes contains a term from a predefined vocabulary of Sinhala legal keywords. All remaining edges receive a default weight of 1.0. These terms were taken from a publicly available website[1] and manually modified to include additional terms in the dataset. These words were validated and discussed in Appendix D.

Given the morphological complexity of Sinhala, POS tagging does not always yield sufficient tagged tokens. When fewer than three content-bearing tokens are produced for a sentence, a fallback mechanism retains all tokens longer than two characters and assigns them a default NOUN tag, preserving graph connectivity and ensuring no sentence is entirely excluded from the graph structure. Figure 1 depicts the word graph generated for a short document.

3.4 Abstractive Candidate Generation

For each cluster, abstractive sentence candidates are generated by performing weighted random walks over the word graph. Starting from the START node, each walk probabilistically selects the next node proportional to edge weights, biasing traversal toward paths containing legal keywords. Each walk produces a candidate of between 8 and 25 words, with up to 8 candidates generated per cluster.

If the random walk procedure yields insufficient candidates due to sparse graphs, a fallback mechanism constructs candidates by ranking nodes by cumulative edge weight and sampling from the top-weighted words. All candidates are then subjected to quality filtering, where sequences shorter than 20 characters, those with less than 60% unique words, or those containing any single word repeated three or more times are discarded.

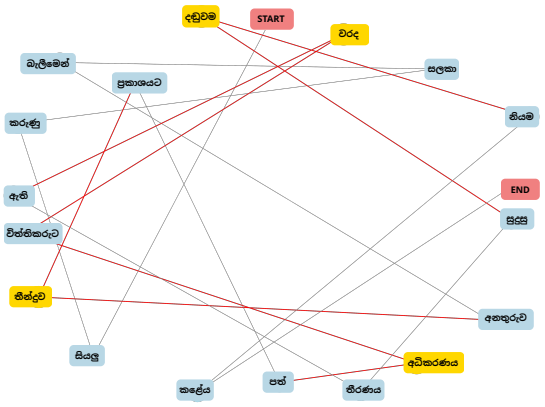

Figure 1: Example of a word graph. START and END nodes are depicted in red. Yellow nodes represent legal keywords, with red edges indicating their connections. Blue nodes represent regular words with grey edges indicating their connections.

3.5 Candidate Scoring

Each abstractive candidate generated from the word graph is evaluated by a weakly supervised sentence scoring model, which assigns a quality score in the range [0,1] reflecting how well the candidate captures meaningful legal content.Candidates are ranked per cluster, and the top two scoring candidates from each cluster are selected for inclusion in the final summary.

3.5.1 Model Selection

Four models were initially evaluated as candidate sentence scorers within the proposed framework: `mBert`, `Llama 3.1 8B`, `Falcon 7B`, and `Laser`. `mBert` was selected as a natural multilingual baseline given its established performance on low-resource languages. `Laser` was included as a lightweight multilingual alternative specifically designed for cross-lingual sentence representation.

`Llama 3.1` 8B and `Falcon 7B` were selected based on perplexity evaluation on the Sinhala legal corpus, where they achieved the lowest perplexity scores of 2.55 and 2.61, respectively, indicating stronger language modelling capability on the target domain compared to other large language models considered (Lasandi and Jayatilleke, 2026).

Based on the summary quality evaluation results

[1] https://talkpal.ai/culture/what-are-the-legal-terms-in-sinhala/

discussed in Section 4, Llama 3.1 8B produced the highest quality summaries among the four models, and therefore, it was selected for continual pre-training (CPT) on the Sinhala legal domain to improve its capacity further. This adds a 5th sentence scorer, the continually pre-trained Llama 3.1, which was then evaluated alongside the original 4 models.

3.5.2 Continual Pre-Training of Llama 3.1

To adapt Llama 3.1 8B to the Sinhala legal domain, continual pre-training was performed on the training section of the SinhaLegal dataset. The training set consisted of 964 legal documents. The resulting domain-adapted model, SinBrief-Legal-Llama 3.1[2], did not surpass the base Llama 3.1 in scoring performance, a finding reported as a negative result. Full implementation details are provided in Appendix A.

3.5.3 Training sentence scorers

A sentence scoring model was trained for each of the five encoder backbones: mBert, Llama 3.1 8B, Falcon 7B, Laser, and CPT Llama 3.1. This was done using a weakly supervised binary classification objective. The full dataset of 1,206 Sinhala legal documents was first partitioned into 80% training (964 documents) and 20% test (242 documents) sets, with the test set held out entirely for final evaluation. The training data was then further split into 80% for model training and 20% for validation using stratified sampling to maintain class balance, yielding 38,920 labelled sentence samples in total for scorer training.

Positive samples (label = 1) were constructed using four strategies applied to the 964 training documents. First, the opening sentence of each document was extracted, as it typically introduces the legal matter at hand. Second, the closing sentence was extracted, as it generally contains the verdict or conclusion. Third, any sentence containing at least one of the 19 predefined Sinhala legal keywords was included. Fourth, sentences exceeding 50 characters were added up to a cap of 5,000 positive samples, as longer sentences tend to carry more substantive legal content.

Negative samples (label = 0) were synthetically generated by applying one of four corruption strategies to the positive sentences: word shuffling, which destroys grammatical structure; random word deletion, which removes approximately 30% of tokens; random word replacement, which substitutes tokens with randomly sampled vocabulary words; and sentence truncation, which retains only the first half of the token sequence. Each positive sentence was corrupted using one randomly selected method, producing an equal number of negative samples and resulting in a balanced dataset. The final training dataset consisted of 38,920 samples in total, which was split 50% positive and 50% negative.

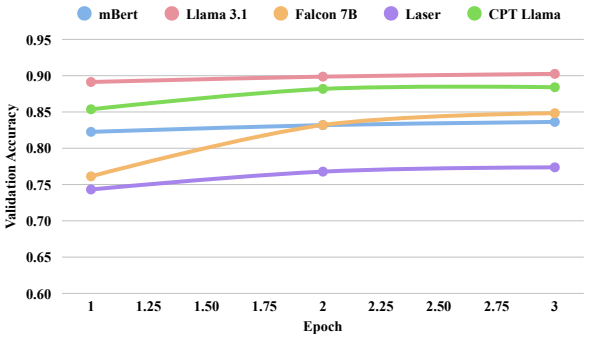

Figure 2: Validation accuracy of the sentence scorers

Each scorer consists of the respective encoder backbone followed by a shared three-layer classification head: a fully connected layer projecting from the encoder's hidden size to 256 dimensions, followed by a second layer reducing to 128 dimensions, and a final layer producing a scalar score in the range [0,1] through a sigmoid activation. ReLU activations and dropout (p=0.3) are applied after each intermediate layer. For mBert, the full model including the backbone was fine-tuned end-to-end, while for the large decoder-only models (Llama 3.1, Falcon 7B, and CPT Llama), only the classification head was trained with the quantised backbone kept frozen, as full fine-tuning was computationally infeasible. For Laser, embeddings were pre-computed using the frozen encoder, and only the classification head was trained, making it the most lightweight of the five configurations.

All models were trained for 3 epochs with a batch size of 16, using the AdamW optimiser with a learning rate of 2×10^{-5}, weight decay of 0.01, and a linear warmup over 100 steps. Binary cross-entropy was used as the loss function, with gradient clipping at 1.0 to ensure training stability.

Figure 2 shows the validation accuracy across epochs for all five scorers. Llama 3.1 achieved the highest validation accuracy of 90.05%, followed closely by CPT Llama at 88.37%. mBert

[2] https://huggingface.co/Minduli-Lasandi/SinBrief-Legal-Llama3.1

and Falcon 7B reached 83.83% and 84.96%, respectively, while Laser achieved 77.25%. The consistently strong performance of Llama 3.1 and CPT Llama suggests that larger decoder-only models may produce more discriminative sentence representations for the Sinhala legal domain. The Training curves of these models are discussed in Appendix B.

3.6 Candidate Selection

Following the scoring stage, all candidates generated for each cluster are ranked in descending order of their scorer output. The top two highest-scoring candidates per cluster are selected for inclusion in the final summary. With three clusters per document, this yields a maximum of six selected sentences. This per-cluster selection strategy ensures that the final summary draws content from semantically distinct regions of the document, preserving topical coverage rather than repeatedly selecting candidates from a single dominant theme.

Qualitative observation of the assigned scores further supports the quantitative findings. Llama 3.1 and CPT Llama assigned scores of 0.9 and above to high-quality legal sentences, demonstrating strong discriminative ability. Laser produced scores of approximately 0.8, and Falcon 7B produced moderate scores of approximately 0.6 and 0.7. mBert assigned the lowest scores overall, with most candidates receiving values near 0.5.

```
Candidate 1: 0.8426 | ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට්පත්‍රයේ...

Candidate 2: 0.6167 | පනතේ වගන්තියේ උපවගන්තියේ ඡේදයෙහි මුද්ල්වල...

Candidate 3: 0.3943 | පනත සහතිකය සටහන්කළේ 1999 අංක දරන විසර්ජන...

Candidate 4: 0.9080 | ශ්‍රීලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ පාර්ලිමේන්තුව...

Candidate 5: 0.0067 | 1998 අංක දරන විසර්ජන පනතේ සිංහල දෙමළ භාෂාපාඨය...

Candidate 6: 0.3943 | පනත සහතිකය සටහන්කළේ 1999 අංක දරන විසර්ජන...

Selected #1 (score: 0.9080)
Selected #2 (score: 0.8426)
```

Figure 3: Scores given by the CPT Llama 3.1.

Figure 3 illustrates an example of candidate scoring by the CPT Llama model, where the two highest-scoring candidates (0.9080 and 0.8426) are correctly selected while lower-quality candidates receive substantially lower scores.

3.7 Final Summary

The top two highest-scoring candidates per cluster are combined to form the final summary. The combined text is then subject to a length control mechanism targeting a character range of 300 to 500 characters, which was determined to be appropriate for a concise legal summary. If it exceeds 500 characters, sentences are added incrementally until the target range is reached, with the final sentence truncated at the nearest word boundary if necessary to stay within the limit. Figure 4 shows an example of a generated summary. The summaries generated by all the 5 models are shown in Appendix C.

ඡන්දය විමසීම පවත් වාගෙන යාම සඳහා යෝජනා පත්‍ර භාරගැනීමේ වේලාවෙන් පැය හතළිස් අටකට පසුව දිනයක් වේලාවක් පාර්ලිමේන්තුව විසින් හින් ඡන්ද පෙත්සම විභාග කිරීම පිළිබඳව තේරීම්. පාර්ලිමේන්තුව විසින් කථානායකවරයාට යවනු ලැබූ ඡන්ද පෙත්සම නිෂ්ප්‍රභ කරමින් ශ්‍රේෂ්ඨාධිකරණයට ඡන්ද අපේක්ෂකයන්ගේ නම් එක් ඡන්දයක් එකතු කිරීමෙන් පසු තමාට පාර්ලිමේන්තුවෙන් ලැබිය හැකි යම් විධානයකට යටත්ව. සිදු වුව තෝරා පත් ගැනීම පුරප්පාඩුව දිනට පසුව හැකි ඉක්මණින් එදින සිට කිසිදු අවස්ථාවක මාස එකක් නොඉක් මවවා පවත්වනු ලැබිය.

Figure 4: Example of a generated summary

4 Evaluation

To evaluate the framework, experiments were conducted on the test set held out. This test set comprised 242 legal documents, representing 20% of the full corpus. The framework was evaluated under five configurations corresponding to each of the five sentence scoring models and compared against five baseline summarisation systems.

4.1 Selection of Evaluation Metrics

Since no reference summaries existed for the SinhaLegal dataset, standard reference-based metrics such as ROUGE cannot be applied. Instead, a set of reference-free evaluation metrics was selected to assess the quality of the generated summaries from multiple perspectives.

Coverage and Density measure the degree to which the summary content is drawn from the source document (Grusky et al., 2018). Coverage is the proportion of summary tokens that appear in contiguous extractive fragments shared with the source. Density is the average squared length of extractive fragments, indicating the length and coherence of copied spans. An abstractive summary is expected to depict lower coverage and density values compared to an extractive summary.

Compression Ratio is defined as the ratio of the summary length in words to the source document length in words (Liu et al., 2022). It quantifies how concisely the framework summarises

the original document. Self-BertScore measures the semantic similarity between the generated summary and the source document using contextual embeddings, without requiring a reference summary (Zhang et al., 2019). Following the approach used for low-resource languages, BertScore is computed using `mBert` as the backbone model.

SummaC is a reference-free factual consistency metric that evaluates whether the content of the summary is entailed by the source document (Laban et al., 2022). It uses a natural language inference model to score sentence-level consistency between the summary and the source, with higher scores indicating greater factual faithfulness. Specifically, SummaCConv internally employs the VitC (VitaminC) NLI model for sentence-level entailment scoring.

4.2 Baseline Systems

Five baseline summarisation systems were evaluated on the test set for comparison with the proposed framework. Three extractive baselines and two abstractive baselines were included. `Lead-3` serves as the simplest extractive baseline, selecting the first three sentences of each document as the summary (See et al., 2017).

`TextRank` (Mihalcea and Tarau, 2004) is a graph-based extractive method that ranks sentences using a PageRank algorithm applied to a sentence similarity graph. Sentence embeddings were computed using *paraphrase-multilingual-MiniLM-L12-v2* to support Sinhala text. `PacSum` (Zheng and Lapata, 2019) extends the `TextRank` approach by introducing directional edge weights in the sentence graph, assigning greater importance to forward-looking sentences.

The fine-tuned `mT5` model[3] is an abstractive baseline fine-tuned on a Sinhala dataset. `BenSumm` (Chowdhury et al., 2021) is an abstractive graph-based summarisation framework originally proposed for Bengali. Since the original model relies on a Bengali-specific ULMFiT backbone, which cannot be applied to Sinhala, `BenSumm` was adapted by replacing the backbone with `mBert`.

Prior to finalising the baseline selection, zero-shot inference was attempted using `mT5` (Xue et al., 2021), and `mBart` (Liu et al., 2020) with no fine-tuning. As shown in Figure 5, both models produced degraded outputs containing special tokens such as *<extra_id_0>*, fragmented Sinhala text, and mixed-script characters, rendering the summaries unusable.

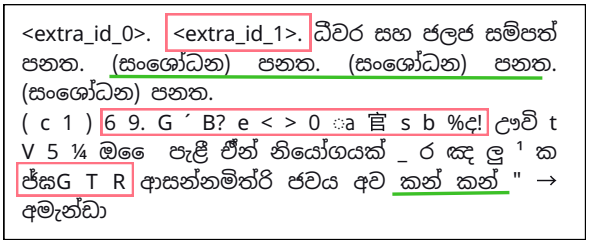

Figure 5: Output from the zero-shot `mT5` method.
*Examples of characters that are not in Sinhala are enclosed in red boxes, while repeated words are highlighted with green underlining.

4.3 Results and Discussion

The results in Table 1 demonstrate that SINBRIEF successfully generates abstractive summaries of Sinhala legal documents, with all five scoring configurations achieving substantially lower Coverage (0.957–0.959) and Density (2.924–3.125) scores than the extractive baselines (Coverage: 1.000, Density: 7.605–8.153). This indicates that the word graph-based candidate generation produces sentences with substantially lower lexical overlap with the source compared to extractive baselines. The fine-tuned `mT5` and adapted `BenSumm` also show reduced Coverage and Density scores (0.855–0.913, 3.989–4.205), though remaining higher than SINBRIEF, indicating that SINBRIEF generates more novel text than any of the baseline systems.

In terms of factual consistency, `Llama 3.1` achieves the highest SummaC score among all evaluated systems (0.752), outperforming not only the other SINBRIEF models but also all five baseline systems, including `TextRank`, which achieves the highest SummaC among baselines (0.648). `Falcon 7B` follows closely (0.740), suggesting that large decoder-only language models provide more effective scoring signals for selecting factually consistent candidates.

Two extractive baselines (`Lead-3` and `PacSum`) achieved unexpectedly low SummaC scores despite directly extracting from the source. This is attributed to the granularity mismatch identified by Laban et al. (2022), where sentence-level NLI models struggle with long, complex sentences that differ significantly from the news corpora on which SummaC was benchmarked (Koreeda and Manning, 2021), with declining reliability further documented for long documents (Mujahid et al.,

[3] https://huggingface.co/Navanjana/Sinhala-Sumarization

Model	Coverage	Density	Comp. Ratio	SummaC	Self-BertScore
Lead-3	1.000	8.153	0.129	0.488	0.850
PacSum	1.000	8.150	0.265	0.481	**0.896**
TextRank	1.000	7.605	0.240	0.648	0.859
Fine-tuned mT5	0.855	3.989	0.099	0.448	0.802
Bensumm	0.913	4.205	0.047	0.452	0.793
Sentence scorers - SINBRIEF					
mBert	0.958	**2.924**	0.084	0.584	0.798
Llama 3.1	**0.956**	3.091	0.089	**0.752**	0.803
Falcon-7B	0.957	2.970	0.088	0.740	0.809
Laser	0.959	3.125	0.084	0.471	0.809
CPT Llama	0.959	3.125	0.090	0.591	0.804

Table 1: Evaluation of existing systems and SINBRIEF sentence scorers using the SINHALEGAL corpus.
Bold: indicates best performance and Underline: indicates the second best. Values shown in the figure are rounded to three decimal places.

2026). As Sinhala legal sentences are particularly long and morphologically complex, the shorter graph-generated candidates of SINBRIEF are more readily verified, and the reported SummaC scores are therefore interpreted as relative indicators.

Regarding the semantic similarity, all SINBRIEF models achieved competitive self-BertScore results (0.798–0.809) compared to the abstractive baselines fine-tuned mT5 (0.802) and BenSumm (0.793), and approached the scores of extractive baselines such as Lead-3 (0.850) and TextRank (0.859). This demonstrates that despite lower lexical overlap with the source, SINBRIEF preserves strong semantic faithfulness to the source document. Among SINBRIEF models, Laser and Falcon 7B achieve the highest Self-BertScore (0.809), while Llama 3.1 emerges as the strongest by achieving the best factual consistency while maintaining competitive abstractiveness and semantic similarity, consistent with its highest validation accuracy of 90.05% during scorer training.

5 Conclusion

This paper presented SINBRIEF, a hybrid abstractive summarisation framework for Sinhala legal documents. The framework combines domain-aware word graph construction with sentence scoring to generate abstractive summaries without requiring annotated training data, addressing a critical gap in NLP resources for Sinhala. Five sentence scoring models were evaluated within the framework, and the evaluation against both extractive and abstractive baselines demonstrated that SINBRIEF produces more abstractive summaries while maintaining semantic faithfulness compared to existing approaches. Furthermore, continual pre-training of Llama 3.1 on the Sinhala legal corpus produced a domain-adapted language model specifically tailored to Sinhala legal text. However, CPT Llama did not surpass the base Llama 3.1 in sentence scoring performance, a finding reported as a negative result.

Future work includes incorporating grammatical constraints into the generation process to improve fluency, conducting human evaluation with native Sinhala speakers, and extending the framework to other low-resource languages.

Limitations

Several limitations of the proposed framework should be acknowledged. The authors, who are native Sinhala speakers, manually identified grammatical errors in the generated summaries, arising from the random walk mechanism which connects words based on co-occurrence weights without grammatical constraints. As a result, the evaluation relies solely on quantitative metrics, highlighting the need for human evaluation in future work.

The stochastic nature of the random walk further means that different summaries may be produced for the same document across runs; due to computational constraints of running large models such as Llama 3.1 and Falcon 7B, multi-seed evaluation was not conducted. The reported validation accuracy of the sentence scorers reflects classification performance rather than ranking quality over graph-generated candidates, which is difficult to evaluate directly in this setting.

Furthermore, the framework relies on Sinling for tokenisation and POS tagging, where frequent errors reduce graph construction quality. The Compression Ratio reflects the fixed length constraint rather than the model's inherent compression abil-

ity, and the individual contributions of domain-aware edge weighting and the legal keyword list remain unquantified due to the absence of ablation studies. Finally, the framework was evaluated on 1,206 Sinhala legal documents, which may not capture the full diversity of legal writing styles.

Acknowledgments

We thank Ms. Sadini Jaburagoda, Attorney-at-Law, for reviewing the Sinhala legal keyword set and providing expert feedback on its relevance to Sri Lankan legal practice. Her evaluation helped assess the legal validity and domain coverage of the selected keywords and provided valuable insights into their applicability across different areas of legal discourse.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Computing Surveys*, 58(6):1–37.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Maryam Azam, Shah Khalid, Sulaiman Almutairi, Hasan Ali Khattak, Abdallah Namoun, Amjad Ali, and Hafiz Syed Muhammad Bilal. 2025. Current trends and advances in extractive text summarization: A comprehensive review. *IEEE Access*, 13:28150–28166.

Florian Boudin and Emmanuel Morin. 2013. Keyphrase extraction for n-best reranking in multi-sentence compression. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 298–305, Atlanta, Georgia. Association for Computational Linguistics.

Veer Chheda, Aaditya Ghaisas, Avantika Sankhe, and Narendra Shekokar. 2025. Extract-explain-abstract: A rhetorical role-driven domain-specific summarisation framework for Indian legal documents. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 439–455, Suzhou, China. Association for Computational Linguistics.

Radia Rayan Chowdhury, Mir Tafseer Nayeem, Tahsin Tasnim Mim, Md. Saifur Rahman Chowdhury, and Taufiqul Jannat. 2021. Unsupervised abstractive summarization of Bengali text documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2612–2619, Online. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International conference on machine learning*, pages 1223–1232. PMLR.

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shibhansh Dohare, Vivek Gupta, and Harish Karnick. 2018. Unsupervised semantic abstractive summarization. In *Proceedings of ACL 2018, Student Research Workshop*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Claire Grover, Ben Hachey, Ian Hughson, and Chris Korycinski. 2003a. Automatic summarisation of legal documents. In *Proceedings of the 9th international conference on Artificial intelligence and law*, pages 243–251.

Claire Grover, Ben Hachey, and Chris Korycinski. 2003b. Summarising legal texts: Sentential tense and argumentative roles. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 33–40.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Ben Hachey and Claire Grover. 2005. Automatic legal text summarisation: experiments with summary structuring. In *Proceedings of the 10th International Conference on Artificial intelligence and Law*, pages 75–84.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Nevidu Jayatilleke and Ruvan Weerasinghe. 2025. A hybrid architecture with efficient fine tuning for abstractive patent document summarization. In *2025 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pages 1–6. IEEE.

Nevidu Jayatilleke, Ruvan Weerasinghe, and Nipuna Senanayake. 2024. Advancements in natural language processing for automatic text summarization. In *2024 4th International Conference on Computer Systems (ICCS)*, pages 74–84.

HMRY Jayawardane. 2022. *Automatic sinhala text summarization for government gazettes using abstractive and extractive methods*. Ph.D. thesis.

Ashish Kasar, Suyash Matade, Durvankur Rasal, and Swapnil Shinde. 2025. Enhancing summarization of legal text documents using pre-trained models. In *2025 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, pages 58–61. IEEE.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Minduli Lasandi and Nevidu Jayatilleke. 2026. SinhaLegal: A benchmark corpus for information extraction and analysis in Sinhala legislative texts. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 114–131, Rabat, Morocco. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. 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.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yizhu Liu, Qi Jia, and Kenny Zhu. 2022. Reference-free summarization evaluation via semantic correlation and compression ratio. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2109–2115, Seattle, United States. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Hemanth Kumar M, Jayanth P, and Anand Kumar M. 2024. Large language models for indian legal text summarisation. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–5.

S Santhana Megala, A Kavitha, and A Marimuthu. 2014. Feature extraction based legal document summarization. *International Journal of Advance Research in Computer Science and Management Studies*, 2(12):346–352.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Zain Muhammad Mujahid, Dustin Wright, and Isabelle Augenstein. 2026. Stress testing factual consistency metrics for long-document summarization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31914–31933, San Diego, California, United States. Association for Computational Linguistics.

Mubashir Munaf, Hammad Afzal, Khawir Mahmood, and Naima Iltaf. 2024. Low resource summarization using pre-trained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(10):1–19.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nikola I Nikolov. 2020. *Abstractive Document Summarization in High and Low Resource Settings*. Ph.D. thesis, ETH Zurich.

Thiago Dal Pont, Federico Galli, Andrea Loreggia, Giuseppe Pisano, Riccardo Rovatti, and Giovanni Sartor. 2023. Legal summarisation through llms: The prodigit project. *arXiv preprint arXiv:2308.04416*.

Juan Ramirez-Orta and Evangelos Milios. 2021. Unsupervised document summarization using pre-trained sentence embeddings and graph centrality. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 110–115, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Jiayu Song, Iman Munire Bilal, Adam Tsakalidis, Rob Procter, and Maria Liakata. 2022. Unsupervised opinion summarisation in the Wasserstein space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8592–8607, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peggy Tang, Kun Hu, Lei Zhang, Junbin Gao, Jiebo Luo, and Zhiyong Wang. 2023. Topiccat: Unsupervised topic-guided co-attention transformer for extreme multimodal summarisation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6643–6652.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jhen-Yi Wu, Ying-Jia Lin, and Hung-Yu Kao. 2022. Unsupervised single document abstractive summarization using semantic units. 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 954–966, Online only. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval*, pages 1949–1952.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Jiawei Zhou and Alexander Rush. 2019. Simple unsupervised summarization by contextual matching. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5101–5106, Florence, Italy. Association for Computational Linguistics.

A Continual Pre-Training of `Llama 3.1`

The documents were tokenised and segmented into fixed-length chunks of 512 tokens using a causal

language modelling objective, where the model is trained to predict the next token given the preceding context.

Given the scale of `Llama 3.1 8B`, full fine-tuning was computationally infeasible. The model was therefore loaded in 4-bit NF4 quantisation via BitsAndBytes, and Parameter-Efficient Fine-Tuning was applied using LoRA. LoRA adapters were attached to all major projection layers, query, key, value, output, gate, up, and down projections, with a rank of r=16 and scaling factor α=32, resulting in a small fraction of trainable parameters relative to the full model size.

Training was conducted for 3 epochs with a batch size of 4, using the AdamW optimiser (Loshchilov and Hutter, 2017) with a learning rate of 2×10^{-5} and a linear warmup over 100 steps. After the completion of the CPT, the LoRA adapter weights were merged back into the base model and saved in bfloat16 precision.

B Training curves

Figure 6 presents the training loss curves for all five sentence scoring models across three epochs. All five models demonstrate a consistent decrease in training loss across epochs, confirming stable convergence without signs of divergence or instability.

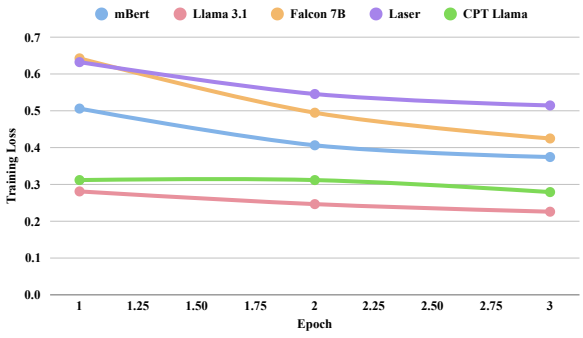

Figure 6: Training loss of the sentence scorers

`Llama 3.1` achieves the lowest training loss across all epochs, starting at 0.281 in epoch 1 and decreasing to 0.2259 by epoch 3, consistent with its strongest validation accuracy of 90.05%. CPT `Llama` follows closely, beginning at 0.3118 and converging to 0.2792, reflecting the benefit of domain adaptation through continual pre-training on Sinhala legal text. `mBert` shows steady convergence from 0.506 to 0.3744, while `Falcon 7B` decreases from 0.6424 to 0.4248. `Laser` records the highest training loss across all epochs, starting at 0.6322 and converging to 0.5145, consistent with its lower validation accuracy of 77.25%.

C Final Summaries

The Table 2 presents the summaries generated by all five SINBRIEF scoring configurations for the same Sinhala legal document. These examples provide a qualitative illustration of how the choice of sentence scoring model influences candidate selection within the same underlying word graph-based generation pipeline. Since all five configurations share the same sentence extraction, clustering, word graph construction, and candidate generation steps, the differences observed across the summaries are attributable to the word graph walks and scoring models' ability to discriminate between high and low-quality candidates.

Despite sharing the same candidate pool, the selected sentences vary across configurations, reflecting the differences in scoring behaviour quantified in Table 2. Models with higher SummaC scores, namely `Llama 3.1` and `Falcon 7B`, are expected to select candidates with stronger factual alignment to the source document, while models with lower discriminative ability, such as `Laser`, may select candidates that are semantically related but less factually grounded. Across all five summaries, grammatical errors, limited fluency, and reduced naturalness of Sinhala text are consistently observable, which is an inherent limitation of the graph-based random walk generation mechanism as discussed in Section 5.

D Legal Keyword Validation

D.1 Word Frequency Analysis

To validate the domain relevance of the selected legal keywords, a corpus-wide frequency analysis was conducted across all 1,206 Sinhala legal documents. For each keyword, the document frequency, the number of documents in which the keyword appears at least once, and the total occurrence count across the entire corpus were recorded. The results, presented in Table 3, confirm that all 19 keywords appear across a substantial portion of the corpus, with high-frequency terms such as පනත \ pʌnʌθʌ, බලය \ bʌlʌjʌ and නීතිය \ ni:θijʌ appearing in over half of all documents. Lower-frequency terms such as පැමිණිලිකරු \ pæmɪnɪlɪkʌrʊ and නීත්‍යනුකූල \ θi:nðʊvʌ appear less frequently but remain relevant to specific legal contexts such as litigation and sentencing.

Model	Generated Summary
mBert	උපවගන්තිය යටතේ හිමිකම් පාන වෙනත් යම් ලිඛිත නීතියක විධිවිධාන වලට අනුකූලව තෝරා පත් ගනු ලබන්නා වූද වෙනත් යම් ලිඛිත නීතියක විධිවිධාන සලස්වා ඇත්තා රීති පනත. කළමනාකාර කමිටුවේ සාමාජිකයන් දෙදෙනා තමන් පැමිණ සිටි බවට ලකුණක් වශයෙන් සාධන පත්‍රයෙහි නම් අත්සන් යුතු. මීගමුව බෞද්ධ සදාචාර වර්ධන සමිතියේ දැනට සාමාජික යනන්ව සිටින තැනැත්තන් මෙයින් සංස්ථාපනය කරනු ලබන සමිතියක් එහි සාමාජිකයන් එකඟ රීති අනුව එකී සමිතියට සම්බන්ධ සියලු කරුණු.
Llama 3.1	1985 අංක දරන මීගමුව බෞද්ධ සදාචාර වර්ධන සමිතිය සංස්ථාගත කිරීම. අත්සන් කිරීමට පරිබාහිර විය යුතු දයක මුදල් පනතේ අයවිය කාර්ය සඳහා සංස්ථාවට ගෙවිය යුතුය. පනතේ වෙනත් යම් ලිඛිත නීතියක විධිවිධාන වලට යටත්ව සංස්ථාවේ පරමාර්ථ ඉටු කිරීම සඳහා අවශ්‍ය බැංකු ගිණුම් විවෘත කිරීමේ පවත්වා ගෙන යාමේ වැසීමේ ඇතිව නැතිව මුදල්. පනතේ වෙනත් යම් ලිඛිත නීතියක විධිවිධාන වලට යටත්ව සංස්ථාවේ පරමාර්ථ ඉටු කිරීම සඳහා අවශ්‍ය බැංකු ගිණුම් විවෘත කිරීමේ කරගෙන යාමේ වැසීමේ ඇතිව නැතිව මුදල් ණයට.
Falcon 7B	මෙහි මින් මතු සමිතිය යනුවෙන් හඳුන්වනු පූර්විකාව ලබන. සංස්ථාවේ රීති වලට පනතෙහි සඳහන් කරනු ලබන සංස්ථාවට මින් මතු සමිතිය යනුවෙන් හඳුන්වනු පූර්විකාව ලබන මීගමුව බෞද්ධ සදාචාර වර්ධන සමිතිය සංස්ථාගත කිරීමේ පවත්වා ගෙන යාමේ. සාමාජිකයන්ගේ ශාරීරික බුද්ධිමය සාමාජික මුදල් සාමාජික ආර්ථික සාධනය නඟාලීම. පනත මීගමුව බෞද්ධ සදාචාර වර්ධන සමිතිය සංස්ථාගත කිරීම කළමනා කාර කමිටුව විසින් කල් තබා අනුමත තිබිය යුතුය.
Laser	පනත මීගමුව බෞද්ධ සදාචාර වර්ධන සමිතියේ කළමනාකාර කමිටුවක් විසින් ගෙවා නිදහස් යුතු යයි මෙයින් ප්‍රකාශ කරනු ලබන සංස්ථාවට මින් මතු සමිතිය සංස්ථාගත කිරීම පරිච්ඡින්න කිරීම. පනතෙහි සඳහන් කරනු ලබන මීගමුව බෞද්ධ සදාචාර වර්ධන සමිතිය සංස්ථාගත කිරීමේ පනත සහතිකය සටහන් කළේ 1985 සැප්තැම්බර් දින ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජයේ ගැසට් පත්‍රයේ. කිරීම බෞද්ධ අධ්‍යාපනය බුද්ධ බෞද්ධයන් පාඨශාලා පිහිටු වීම නඩත්තු ප්‍රවර්ධනය ධර්මය ප්‍රචාරණය අතර. පිහිටු බෞද්ධයන් පාඨශාලා බුද්ධ අධ්‍යාපනය බෞද්ධ කිරීම.
CPT Llama	ඒවා සංශෝධනය කිරීම සාක්ෂිකරුවකු වශයෙන් යම් තැනැත්තකුගේ අත්සන් කිරීමට පරිබාහිර විය යුතු අතර සමගිය ඇති කිරීම සාක්ෂිකරුවකු වශයෙන් සාධන පත්‍රයකට නොතැබිය යුතු දයක මුදල් පනතේ. බුද්ධ ධර්මය ප්‍රචාරණය කිරීම පරිච්ඡින්න කිරීම කළමනා කාර කමිටුව විසින් ගෙවා නිදහස් යුතු. පනතේ වෙනත් යම් ලිඛිත නීතියක විධිවිධාන වලට යටත්ව සංස්ථාවේ පරමාර්ථ ඉටු කිරීම සඳහා අවශ්‍ය බැංකු ගිණුම් විවෘත කිරීමේ කරගෙන යාමේ වැසීමේ ඇතිව නැතිව මුදල් ණයට.

Table 2: Summaries generated by all five SINBRIEF models for the same source document. Document used was *acts_1985-09-17_meegamuwa_bauddha_sadachara_204* taken from the `SinhaLegal` Dataset.

As observed in Table 3, all 19 keywords record a total occurrence count exceeding 50 across the corpus, with the exception of පැමිණිලිකරු \ pæmɪnɪlɪkʌru which records 39 total occurrences. Despite its relatively lower frequency, this term remains a legally significant keyword specific to litigation contexts. The remaining 18 keywords collectively account for thousands of occurrences with පනත \ pʌnʌθʌ alone recording 24,293 occurrences across all 1,206 documents.

The keyword list was initially sourced from a publicly available Sinhala legal terminology reference[4] and subsequently extended through manual inspection of the corpus to include additional domain-specific terms. The corpus frequency analysis presented in Table 3 further confirms the domain relevance of the selected keywords, demonstrating that all 19 terms are genuinely embedded in Sinhala legal writing across the corpus.

Keyword	Doc. Frequency	Occurrences
පනත \ pʌnʌθʌ	1206	24293
බලය \ bʌlʌjʌ	773	3982
නීතිය \ ni:θijʌ	715	2733
සාක්ෂි \ sa:kʃɪ	487	1409
අධිකරණය \ ʌðʰɪkʌrʌnʌjʌ	354	2138
කොන්දේසි \ kɒnðe:sɪ	321	912
වරද \ vʌrʌðʌ	299	4167
ව්‍යවස්ථා \ vyʌvʌsθʰa:	257	1304
නඩුව \ nʌduvʌ	214	830
දඬුවම \ ðʌduvʌmʌ	183	455
අභියාචනය \ ʌbʰɪja:tʃʌnʌjʌ	130	698
විනිශ්චය \ vɪnɪʒtʃʌjʌ	119	715
දෝෂ \ ðɔ:ʃʌ	99	121
චෝදනා \ tʃɔ:ðʌnʌ:	72	225
ප්‍රතිපාදන \ pkrʌθɪpa:ðʌnʌ	52	80
කොන්ත්‍රාත්තුව \ kɒnθra:θθuvʌ	46	51
පැමිණිල්ල \ pæmɪnɪllʌ	37	83
නීන්දුව \ θi:nðuvʌ	33	79
පැමිණිලිකරු \ pæmɪnɪlɪkʌru	10	39

Table 3: Corpus frequency analysis of Sinhala legal keywords across 1,206 documents.

D.2 Human Validation

To further validate the selected keywords, they were reviewed by a legal domain expert with professional experience in Sri Lankan legal practice. The objective of this evaluation was to assess whether the keywords represent meaningful legal concepts commonly used in Sinhala legal discourse.

The expert confirmed the relevance of the majority of the selected keywords and their association with legal concepts. The review also identified two keywords, බලය \ bʌlʌjʌ and කොන්දේසි \ kɒnðe:sɪ, as being more strongly associated with constitutional contexts than with the broader body of legal practice. While alternative terms were suggested, these keywords were retained because the corpus frequency analysis demonstrated their substantial presence across the legal document collection.

Overall, this expert review and validation support the legal relevance of the selected keywords, while the corpus frequency analysis provides quantitative evidence of their usage throughout the

[4] https://talkpal.ai/culture/what-are-the-legal-terms-in-sinhala/

dataset. Together, these validation methods indicate that the selected keywords represents Sinhala legal text and suitable for legal-domain adaptation.