SinMix2Mono: A Dataset for Code-mixed Romanized Sinhala Translation and Transliteration

Rukshan Dias[1] **, Deshan Sumanathilaka**[2] **, Archchana Sindhujan**[3] **, Minidu Nimna**[1]
[1]Informatics Institute of Technology, Colombo 06, Sri Lanka
[2]Swansea University, Wales, UK
[3]Surrey University, Guildford, UK
rukshan.20210046@iit.ac.lk, t.g.d.sumanathilaka@swansea.ac.uk
a.sindhujan@surrey.ac.uk, minidu.20210462@iit.ac.lk

Abstract

Code-mixed and Romanized texts are widely used in digital content, yet they remain largely underexplored for many low-resource languages, including Sinhala. The scarcity of high-quality parallel data has limited progress on downstream tasks, such as machine translation and transliteration. We introduce **SinMix2Mono**, the largest manually annotated parallel training dataset, followed by the first gold standard benchmark and code-mixed transliteration ambiguity corpora for code-mixed romanized Sinhala to Sinhala conversion. The dataset comprises approximately 25,000 real-world sentences collected from social media, covering diverse domains and authentic code-mixing patterns. To ensure high-quality translations, we used an annotation pipeline that combined rule-based transliteration, LLM-assisted translation, and human validation. The golden test dataset, which includes 2549 sentences, and the code-mixed transliteration ambiguity test were validated by three annotators, yielding Gwet's AC1 scores of 0.7465 and 0.7068, respectively. We benchmarked nine systems, including statistical, neural and commercial LLMs. SinMix2Mono[1] provides a robust training and evaluation resource, establishing a strong benchmark for future research on Sinhala code-mixed translation and transliteration.

[1]SinMix2Mono Dataset

1 Introduction

With the advent of Web 2.0, user participation, social media engagement, and cloud-based technologies have experienced substantial growth. At the same time, digital platforms have become more accessible for everyday use through improved support for native languages (Mudiyanselage and Sumanathilaka, 2024). As a result, users increasingly prefer to communicate in languages they are most comfortable with, either in their native scripts or in romanized forms (Shibli and others, 2023). This trend has led to greater user engagement on digital platforms, particularly among speakers of low-resource languages (Dharmasiri and Sumanathilaka, 2024).

Code-mixing refers to the mixing of different languages in communication, which has increased with globalization. Romanized text refers to writing any particular language with Latin script format using phonetic typing rather than using Unicode, which has been identified as more convenient among bilingual and multilingual speakers (Kugathasan and Sumathipala, 2020). Both code-mix and romanization would be considered as noisy text, considering their non-standard format. Machine translation is useful not only for translating one standard language into another but also for translating noisy data into a more standardized format of script. This would add more control and value to the data that is being used. Allowing for the processing of noisy data for different linguistic tasks. Sinhala is spoken and used by 17 million people in the Sri Lankan community (Ranasinghe et al., 2025). Sinhala being a low-resource language, code-mixed romanized Sinhala is considered a niche research area, which has a very limited number of existing studies.

The scarcity of available data is one of the primary factors contributing to the limited number of studies on code-mixed, romanized Sinhala. Even though there are a few datasets on sentiment analysis and language identification on code-mixed romanized Sinhala (Uthpala and Thirukumaran, 2024; Rathnayake et al., 2022; Smith and Thayasivam, 2019), there are extremely limited parallel data with code-mixed romanized Sinhala and its relevant Sinhala translation or transliteration. Despite the importance of the problem, there are very limited dedicated resources that could support or evaluate machine translation systems.

To the best of our knowledge, this work introduces the largest parallel training dataset of code-mixed romanized Sinhala paired with corresponding Sinhala translations. Furthermore, we present the first gold-standard dataset and an ambiguity dataset designed for the evaluation of code-mixed romanized Sinhala. In addition to constructing the dataset, we benchmarked across 9 existing models that support back-transliteration. The evaluation has highlighted how accurately these models would handle different code-mixing lexical patterns and word ambiguity.

The contributions of this study are as follows:

- Introducing the largest parallel dataset for Code-mixed romanized Sinhala translation and transliteration.
- Introducing the first manually annotated Golden dataset on Code-mixed romanized Sinhala, and code-mixed transliteration ambiguity dataset.
- Benchmarking the Golden dataset on 9 existing models that support back-transliteration.
- Publicly releasing the dataset for the use of future studies on the domain of Sinhala Machine Translation.

In summary, this introduction has outlined the research landscape and highlighted the key themes addressed in this work. The subsequent sections review related studies, describe the proposed methodology for dataset creation, present and analyze the results of current state-of-the-art performance systems for Sinhala, and discuss the limitations of the approach. We also identify directions for future research.

2 Related Work

This section reviews publicly available datasets for code-mixed and romanized translations in Sinhala and other low-resource languages. Despite growing interest in this area, publicly available datasets for Sinhala remain scarce. As a result, we primarily examine existing transliteration datasets developed for Sinhala and comparable low-resource languages.

2.1 Sinhala Transliteration

Recent studies associated with the SwaBhasha series have made substantial contributions to Sinhala machine transliteration (Sumanathilaka et al., 2025b). In particular, the dataset introduced by Sumanathilaka et al. (2025b) comprises 7.2 million manually curated ad-hoc transliterations from Romanized Sinhala to native Sinhala script. A subsequent study further introduced a 7-million sentence-level dataset generated through automatic mapping techniques (Sumanathilaka et al., 2025b). In addition, several other recently released datasets have focused on evaluating transliteration ambiguity and general machine transliteration in Sinhala (Perera and Sumanathilaka, 2025; Sumanathilaka et al., 2025a). However, all of these resources are limited to Romanized Sinhala-to-Sinhala transliteration and do not address broader code-mixed or translation settings.

Despite these improvements, the dataset created by (Kugathasan and Sumathipala, 2021) remains the only existing dataset on code-mixed romanized Sinhala. It contains code-mixed romanized Sinhala and its relevant native Sinhala translation, which makes it suitable for machine translation tasks. The dataset consists of 5,000 parallel sentences, where the code-mixed romanized Sinhala sentences were collected from social media platforms. The collected sentences have been manually translated into Sinhala with the help of a human translator. The translated dataset has been validated using the crowdsourcing method. The study also proposes an LSTM-based machine translation model for code-mixed romanized Sinhala to Sinhala, which has been trained on this dataset.

2.2 Non Sinhala Transliteration

The *COMI-LINGUA dataset* (Sheth et al., 2025) is considered the largest manually annotated Hindi parallel corpus. It includes 24,558 parallel sentences that contain romanized Hindi text with the relevant native Hindi translation. The data collection has been conducted on diverse platforms and domains, including news, social media, politics, and online archives. This dataset is designed for machine translation benchmarking.

The *PHINC dataset* (Srivastava and Singh, 2020) contains code-mixed romanized Hindi text with Hindi translation. The dataset comprises data collected from social media platforms, primarily from Twitter. A total of 13,738 parallel sentences are present in this dataset. Although it's less in terms of quantity compared to the COMI-LINGUA dataset (Sheth et al., 2025), the PHINC dataset is well-established and widely used in machine translation shared tasks.

The study by Wisal et al. (2022) has attempted to construct a dataset on code-mixed romanized Urdu to Urdu translation, considering the inaccessibility of existing relevant data. Multiple annotators have been included in this task, and they have initially translated 20,000 sentences. The pre-processing techniques, such as removing extra columns, duplicate sentences, and unavailable translations, have been followed by the annotators. The final revised dataset consists of 17,689 parallel sentences, which have been divided into 80% as a training set and 20% as a testing set.

The *HinGe dataset* (Srivastava and Singh, 2021) contains Hindi-English code-mixed data, compiled to address the scarcity of high-quality data on the language group. The data in the dataset could be structured into two components: Human-generated and Machine-generated sentences. The human-generated section contains 4,803 sentences, which have been annotated by five expert annotators. The machine-generated data have been synthetically generated using Word-align Code-mixing (WAC) and Phrase-align Code-mixing (PAC) algorithms. Authors have mentioned that, in addition to machine translation, this corpus can also be used for other NLP tasks, such as language identification and POS tagging.

3 Dataset Creation

This section of the paper describes the development process of the SinMix2Mono dataset. This includes the aim of this creation, the selection of code-mixed romanized Sinhala sentences, sentence annotation, and the formulation of the final dataset. Based on the findings from related works, it is evident that existing datasets are not well-suited for machine translation and transliteration tasks involving code-mixed romanized Sinhala text, primarily due to the limitation of a quality parallel data corpus. To address this challenge, we introduce the SinMix2Mono dataset, which is specifically designed to support the training and evaluation of Sinhala romanized code-mixing systems. The dataset aims to facilitate systematic research on both translation and transliteration in Sinhala-English code-mixed settings by providing adequately annotated and representative data.

3.1 Data Collection

The study by Kumbhar and Thakre (2024) highlights that informal communication patterns, such as romanized code-mixing, are prevalent on digital social media platforms, including Facebook, YouTube, and WhatsApp. Motivated by this observation, we construct our dataset using user-generated content from publicly available YouTube comments, which frequently exhibit code-mixed romanized Sinhala. To this end, we developed an algorithm to automatically collect comments containing romanized Sinhala code-mixing.

To ensure domain diversity, data were collected across the following categories: News and Law/Politics, Science and Technology, Business and Education, Entertainment, and General content. The General category comprises comments from podcasts and interviews covering a wide range of topics. In total, approximately 20,000 comments were collected across all categories. Figure 2 illustrates, and Table 3 summarizes the distribution of the collected data among these domains.

In addition to the scraped data, we incorporate instances from the Swa-Bhasha dataset (Sumanathilaka et al., 2024), which provides parallel Romanized Sinhala-Sinhala data. Upon analysis, we observed that this dataset also contains a subset of code-mixed sentences. Accordingly, we applied the same filtering algorithm to extract sentences that consist of code-mixed romanized Sinhala. By combining data collected from existing resources with content scraped from real-world digital platforms, the resulting dataset achieves greater linguistic diversity and better reflects authentic usage patterns observed in natural communication.

3.2 Data Preprocessing

Following data collection from YouTube comments and the extraction of code-mixed romanized Sinhala from existing datasets, the aggregated data undergo several preprocessing steps. These in-

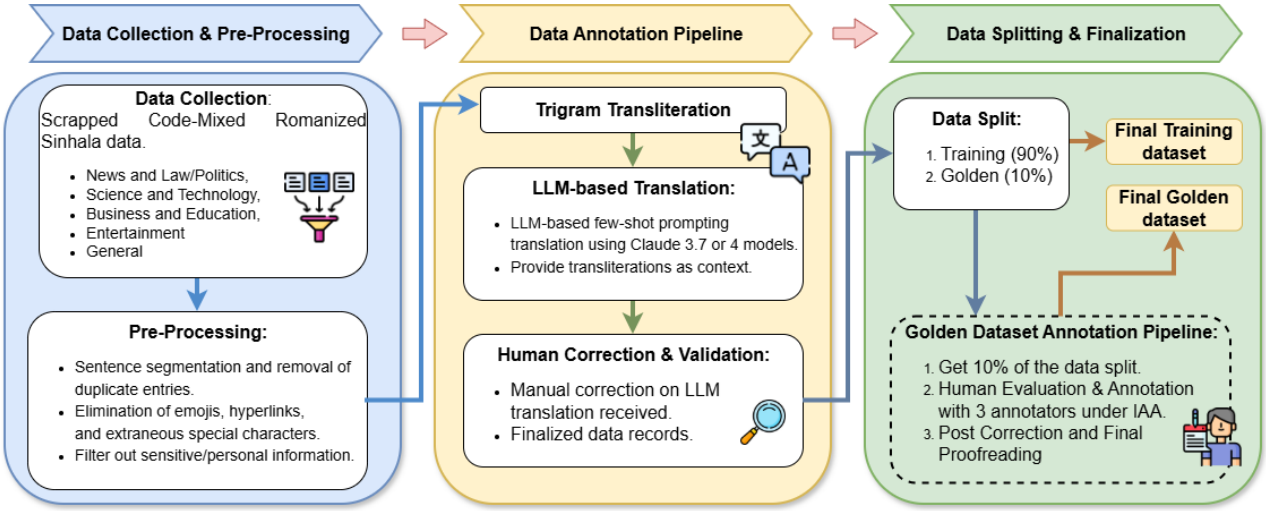

Figure 1: SinMix2Mono dataset creation and annotation workflow.

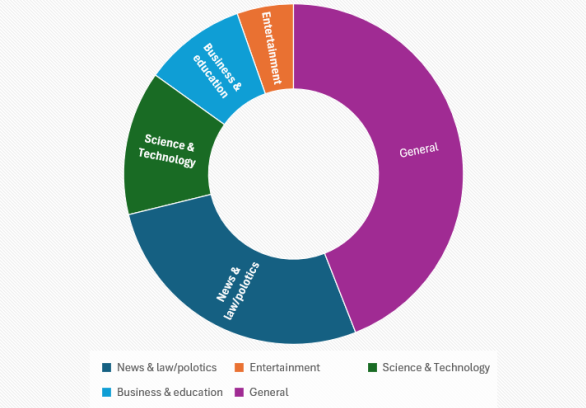

Figure 2: Domain distribution of scraped data. The general category contains comments that are not classified into any of the above areas, including podcasts and interviews.

clude sentence segmentation, removal of duplicate entries, and elimination of emojis, hyperlinks, and extraneous special characters. We removed usernames, phone numbers, addresses, and other personal information from the sentences to anonymize user identifiers to protect privacy. This preprocessing stage is essential to reduce noise and to ensure that the final dataset predominantly consists of high-quality code-mixed romanized Sinhala instances.

3.3 Data Annotation

Following preprocessing, candidate sentences for annotation were further filtered to ensure high data quality. As the first step in the annotation pipeline, the collected data was transliterated using the trigram-based and rule-based model proposed by Sumanathilaka et al. (2023), which produces Sinhala Unicode output and handles ad hoc transliteration patterns. However, due to limitations of the transliteration model, this step is limited to transliteration and English words embedded within Romanized Sinhala were not translated into native Sinhala.

Consequently, a second phase of the annotation required translating the English words into their appropriate Sinhala equivalents based on sentence-level context. For this purpose, Claude-Sonnet-3.7[2] and Claude-Sonnet-4[3] were employed using a carefully designed instruction prompt. The prompt incorporates a few-shot learning strategy, providing both positive and negative examples to guide correct translation and transliteration of code-mixed Romanized Sinhala. These models were selected due to their demonstrated support for Sinhala and their effectiveness in low-resource translation tasks with few-shot prompting (Jayakody and Dias, 2024).

The prompt, provided in Appendix D, supplies the LLM with both the original code-mixed Romanized Sinhala input and its trigram-based transliteration as context information.

While the model generated reasonably fluent Sinhala translations, several issues were observed, as listed below with the severity.

- Ambiguity - (major): LLM failed to perform translation and transliteration for ambiguous words, producing incorrect translations.
- Selection of translation & transliteration - (major): LLM performed translations, where transliteration is expected. Eg: Named entities

[2]https://www.anthropic.com/news/claude-3-7-sonnet
[3]https://www.anthropic.com/news/claude-4

are expected to be transliterated, but LLM perform word-level translation, which is incorrect.

- Unsuitable translations - (moderate): For certain English words, LLM have produced more formal translations, which are correct on the word level. However, because the collected data are informal in nature, formal translations of this do not provide an accurate contextual interpretation.
- Sentence restructuring - (minor): LLM did produce translations that contained major restructure of the sentence, with some enhancement with additional words.

As a result, human post-editing and validation were necessary to ensure translation accuracy and consistency. Annotators manually reviewed each sentence, correcting erroneous or mismatched transliterations and translations. This human-in-the-loop annotation phase was the most resource-intensive stage of dataset creation, requiring substantial human effort and computational resources. Due to the limited availability of annotators, the training set was only validated by the primary author of the work.

4 Gold Dataset Creation process

As discussed in the previous sections, the data collection, preprocessing, annotation, and validation pipeline resulted in a high-quality corpus comprising nearly 25,000 code-mixed Romanized Sinhala sentences paired with their corresponding Sinhala translations. From this corpus, we constructed a gold-standard evaluation set intended for benchmarking machine translation models on the code-mixed Romanized Sinhala-Sinhala translation and transliteration task. The release of this gold dataset is expected to support and standardize future research on Sinhala code-mixing and low-resource translation.

To create the gold dataset, 2,549 sentences (approximately 10% of the full corpus) were randomly sampled, with the additional constraint that each sentence contained between 4 and 20 words. Inter-annotator agreement (IAA) was conducted to assess annotation quality and validity. Three native Sinhala speakers served as annotators, each independently evaluating the code-mixed Romanized Sinhala input and its corresponding Sinhala translation. Detailed annotation guidelines were provided to ensure consistency across annotators. Each instance was labeled using one of three categories: *Correct*, *Incorrect*, or *Invalid*.

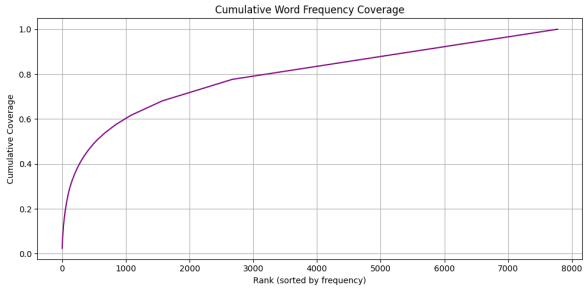

Figure 3: Cumulative word frequency

A sentence pair was labeled *Correct* if the source and target were semantically and linguistically aligned according to the guidelines. The *Incorrect* label was assigned when the translation was inaccurate or incomplete, while *Invalid* was used for instances outside the scope of the task, such as fully English sentences or sentences not satisfying the 4-20 words constraint. Upon completion of annotation, agreement was measured using Gwet's AC1 (Gwet, 2008), yielding a score of 0.7465 with a raw agreement of 71.78%. Gwet's AC1 was preferred over Fleiss' Kappa due to Kappa's known limitations under high chance agreement and its lack of robustness to imbalanced category (Feinstein and Cicchetti, 1990; Gwet, 2008).

Following the agreement analysis, all instances labeled as *Incorrect* were reevaluated and corrected, while *Invalid* instances were removed from the dataset. Sentences labeled as *Correct* were retained without modification. In addition, qualitative feedback from annotators revealed recurrent use of certain English lexical items across multiple instances. To improve lexical diversity, sentences containing the same code-mixed words were replaced with newly curated, unique code-mixed Romanized Sinhala sentences. These refinements further enhance the diversity and reliability of the gold dataset, making it suitable for robust evaluation of machine translation. Statistics of both training and golden parallel corpus are listed in Table 1.

Figures 3 and 4 present characteristics of the word frequency on the golden dataset using Zipf's Law (2) (Zipf, 1949) and cumulative word frequency coverage (1). The rank-frequency distribution shows an approximately linear decay, indicating a Zipfian distribution where a small number of high-frequency words dominate the corpus, while a large number of low-frequency words form a long tail. This behavior is characteristic of lan-

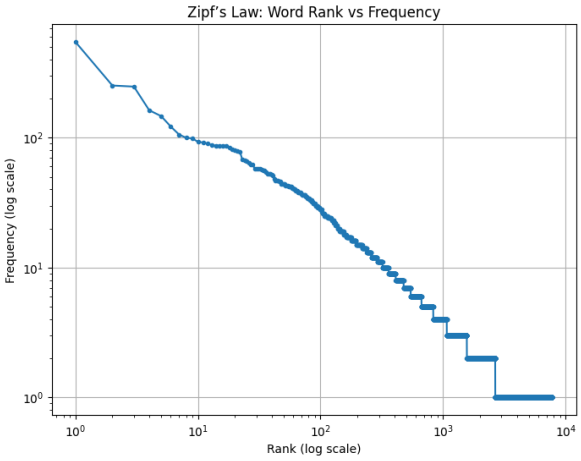

Figure 4: Zipf's Law Word Rank Frequency distribution

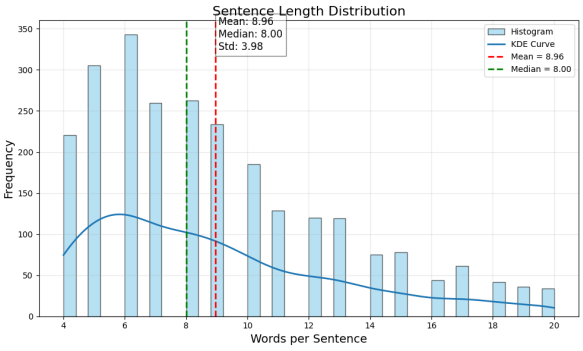

Figure 5: Sentence length distribution

guage used on social media and is particularly common in code-mixed text.

The cumulative coverage analysis further shows that a limited subset of frequent tokens accounts for a substantial proportion of total word occurrences, while thousands of additional tokens contribute incrementally toward full coverage. The gradual saturation of the curve reflects substantial lexical diversity, likely arising from informal usage, spelling variation, and code-mixing. These results indicate that the dataset contains realistic code-mixed linguistic properties and is not dominated by artificial repetition or excessive noise.

$$C(k) = \frac{\sum_{i=1}^{k} f_i}{\sum_{j=1}^{N} f_j} \quad (1)$$

$$f(k) = \frac{C}{k^s} \quad (2)$$

The Figure 5 visualizes the sentence length distribution in the dataset. The structural analysis of the corpus reveals distributional and correlational characteristics. The sentence length distribution, truncated to a range of 4 to 20 words, exhibits a Positively Skewed distribution, where

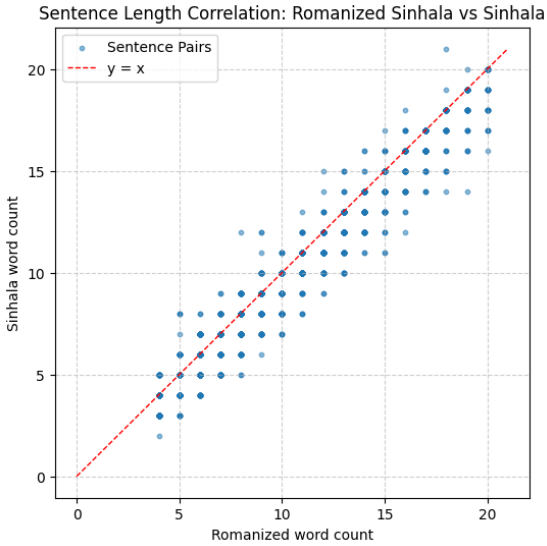

Figure 6: Romanized Sinhala, Sinhala sentence length correlation

the mean length (8.96) exceeds the median (8.00). This indicates that the dataset primarily consists of short, simple sentences, as commonly used in social media. Furthermore, in Figure 6, the alignment between Romanized and native Sinhala sentence lengths demonstrates a strong positive linear correlation closely following the identity line, yet closer to the x-axis. This indicates that the romanized version of Sinhala often contains more words than native Sinhala. This is expected for code-mixed text, given its lexical patterns.

5 Code-mixed Transliteration Ambiguity Dataset Creation Process

In addition to the gold-standard dataset described in the previous section, we compiled a separate dataset to evaluate ambiguous words in code-mixed Romanized Sinhala. Initially, we selected 34 ambiguous code-mixed transliterated words using the LTRL frequency-mapped dataset with dual valid meanings. Sentences were then synthetically generated using Gemini 3 Flash[4] and manually corrected for clarity. For each word, 15 sentences were generated: five sentences representing the Sinhala sense, five representing the English sense, and five containing both senses combined. Using this approach, 15 sentences with relevant translations were generated for each word, yielding a total of 510 data records. The selected words and sample generated sentences are presented in Figure 12 and Figure 13.

[4]https://deepmind.google/models/gemini/flash/

Similar to the initial golden dataset, an IAA-based evaluation and annotation were conducted for the code-mixed transliteration ambiguity dataset to ensure its quality and validity. Three native Sinhala annotators with English proficiency labeled each generated sentence and its translation as either correct, incorrect, or invalid. Records marked as incorrect or invalid were subsequently reviewed and corrected by the first author. The same IAA procedure was applied to this dataset evaluation, resulting in a Gwet's AC1 score of 0.7068.

6 Dataset Evaluation

To establish the applicability of the proposed Golden dataset, we have conducted a comparative evaluation across 9 existing models. The evaluation has focused on three main metrics relevant to machine translation: BLEU (Papineni et al., 2002), chrF++ (Popović, 2015), and BERTScore (Zhang et al., 2019).

6.1 Systems Evaluated

To evaluate the effectiveness of the Golden Dataset, we have assessed nine existing models mentioned below. These systems vary from rule-based to Neural and hybrid models and to state-of-the-art LLMs, offering wider quality comparison.

- **Rule-Based**: The system implemented by UCSC (2006) has been used as the rule-based system for this evaluation. The system is capable of transliterating romanized Sinhala to Sinhala Unicode by defining a set of rules and mappings, but not capable of handling ad-hoc typing patterns and code-mixed text.
- **Swa-Bhasha trigram**: The system proposed by Sumanathilaka et al. (2023) follows a hybrid approach to transliterate Romanized Sinhala to the Sinhala native script. The architecture uses a trigram-based statistical model and a rule-based transliterator. The study focuses on transliterating non-standard and ad-hoc romanized Sinhala content, while considering code-mixing translation as out of scope.
- **Sinhala-Bert and finetuned model**: The Sinhala-Bert model (Ransaka, 2023) and its fine-tuned model, implemented by Perera et al. (2025), have been considered for this evaluation task. Similar to the Swa-Bhasha trigram model, these models consider code-mixed translation as out of scope. This study (Perera et al., 2025) reveals that the finetuned model outperforms Swa-Bhasha (Sumanathilaka et al., 2023) in terms of transliteration quality.
- **Qwen-3 Max**: The Qwen-3 Max[5] model was selected for evaluation due to its strong multilingual capabilities, supporting 119 languages, including Sinhala. The model is known for its robust cross-lingual representations and has demonstrated competitive performance on machine translation tasks, making it well-suited for evaluating code-mixed Romanized Sinhala-Sinhala translation. (Team, 2025).
- **Gemini 2.5 flash-lite**: The Gemini 2.5 flash-lite[6] model contains advanced reasoning and contextual understanding abilities, which allow it to handle the word ambiguity challenge in code-mixing. A study conducted by Bhattacharjee et al. (2024) has found that Gemini's performance on Code-mixed Hindi-English translations outperforms other models (GPT-4, BLOOMZ-3B, GEMMA).
- **XLM-RoBERTa model**: The XLM-RoBERTa model[7] is an encoder-only transformer and the multilingual version of RoBERTa. For this task, this model has been fine-tuned by wrapping two encoder models inside an EncoderDecoderModel, making it suitable for a translation task.
- **M2M100 model**: The m2m100 418M[8]model is a multilingual encoder-decoder (seq-to-seq) model trained for Many-to-Many multilingual translation. This includes the Sinhala language with id 'si'.
- **Swa-Bhasha mBART model**: The Swa-Bhasha mBART model[9] is a finetuned version of MBart, that is designed to transliterate romanized Sinhala text to native Sinhala. The main limitation of this model is that it would not address code-mixing.

For evaluation purposes, the XLM-RoBERTa, m2m100, and Swa-Bhasha mBART models were fine-tuned on the SinMix2Mono training set. All three models were trained for 5 epochs using Native Automatic Mixed Precision (AMP) and the AdamW optimizer to balance efficiency and performance. The XLM-R and SwaBhasha mBART

[5]https://qwen.ai/blog?id=qwen3-max
[6]https://deepmind.google/models/gemini/flash-lite/
[7]https://huggingface.co/FacebookAI/xlm-roberta-base
[8]https://huggingface.co/facebook/m2m100_418M
[9]https://huggingface.co/deshanksuman/swabhashambart50SinhalaTransliteration

Split	Type	#Sent.	#Tokens	Vocab
Train	Roman	22,225	201,228	33,434
	Sinhala	22,225	193,863	25,555
Golden	Roman	2,549	22,833	7,999
	Sinhala	2,549	21,836	6,726
Ambiguity	Roman	510	4,605	1,289
	Sinhala	510	4,321	1,475

Table 1: Statistics of training, golden, and ambiguity datasets.

models were fine-tuned using Low-Rank Adaptation (LoRA) ($r = 128, \alpha = 256$), while M2M100 underwent Supervised Fine-Tuning. Specifically, XLM-R utilized a linear learning rate scheduler with a total batch size of 64, while M2M100 followed a similar linear path with a more conservative learning rate of 2×10^{-5} and a batch size of 32. In contrast, the swaBhasha mBART model employed a cosine scheduler with a 5% warmup phase to gradually adjust its 1×10^{-4} learning rate.

6.2 Evaluation Procedure

All nine selected models were evaluated using the full gold-standard dataset. For the translation-specific models, namely the rule-based system, Swa-Bhasha trigram model, Sinhala-BERT, and fine-tuned Sinhala-BERT, each code-mixed Romanised Sinhala sentence was provided as input, with the expectation of producing a Sinhala Unicode translation as output. The generated translations were then compared against the corresponding human-annotated Sinhala references in the golden dataset.

For the LLMs (Qwen-3 Max and Gemini 2.5 Flash-Lite), evaluation was conducted under a zero-shot and few-shot prompting setting. The temperature was maintained at zero to ensure reproducibility. The following instruction prompt was used for all inputs in zero-shot evaluation.

Prompt for zero-shot evaluation.

You are an expert in Translation and Transliteration. Translate/transliterate the input Code-Mixed Romanized Sinhala sentences into natural monolingual Sinhala. Transliterate all Romanized Sinhala words into Sinhala while translating any English words in the sentence to Sinhala. Only return the Sinhala context as output with the SAME ORDER. Do not return anything else.
Input: original_sentence.
Output: "

A few-shot evaluation was conducted with 3 samples with the prompt provided in Figure 15. Similar to the task-specific translation models, the LLMs generated translated and transliterated Sinhala outputs, which were subsequently evaluated against the gold-standard references using multiple automatic evaluation metrics.

6.3 Evaluation Metrics

This section would describe the selected evaluation metrics that have been used for this code-mixed romanized Sinhala translation and transliteration evaluation task.

BLEU is considered the most widely used evaluation metric for machine translation tasks. BLEU calculation would measure the precision of n-grams in candidate text against reference text with Brevity Penalty to overcome short translations (Papineni et al., 2002). For this evaluation task, the **SacreBLEU** metric has been used to ensure reproducible and comparable BLEU scores, since BLEU is highly sensitive to pre-processing choices (Post, 2018).

chrF++ combines the character-level matching with the lexical accuracy of word-level matching (Popović, 2017). Since code-mixed and transliterated outputs exhibit different character-level variations, chrF++ captures them much better than BLEU.

BERTScore , unlike traditional n-gram metrics, leverages contextual embeddings from pretrained models to measure the similarity between the hypothesis and the reference (Zhang et al., 2019). By capturing semantic similarity, this metric preserves meaning even when translations are phrased differently which is a common challenge when translating code-mixed tokens.

7 Results and Analysis

This section of the paper discusses the evaluation results of 9 models against the proposed golden dataset. Each system has been evaluated with 3 evaluation metrics. The results are summarised in Table 5 and visualized in Figure 7.

7.1 Observations

The NMT models trained on the SinMix2Mono corpus achieved the highest overall performance, consistently outperforming all other models across

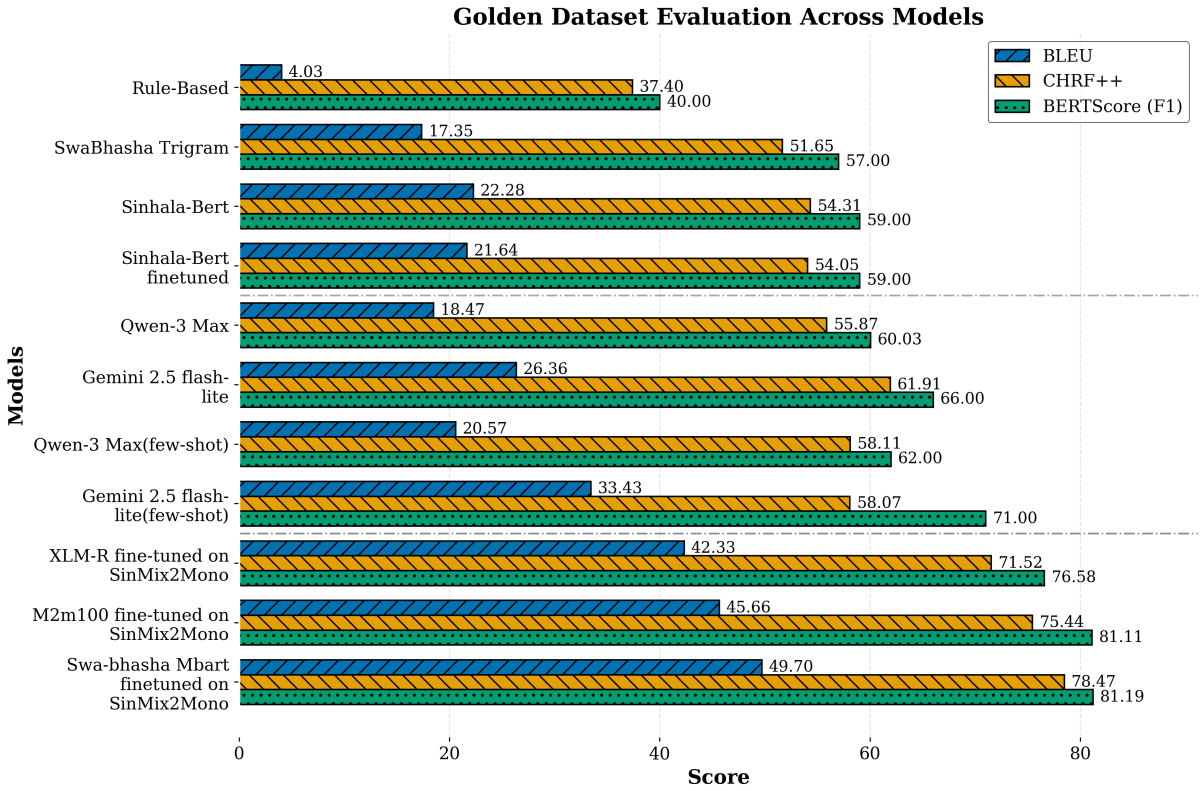

Figure 7: Model evaluation on SinMix2Mono golden dataset

all metrics. Both Gemini 2.5 Flash-lite and Qwen-2.5-Max demonstrated competitive results, showing slightly improved performance in both few-shot and zero-shot settings. The Sinhala-BERT base and fine-tuned models produced nearly identical translation outcomes.

In contrast, the SwaBhasha trigram model and the pure rule-based model yielded the lowest scores. These results highlight the inherent limitations of statistical and rule-based approaches in handling the non-standardized nature of romanized code-mixed data.

7.2 Error Analysis and Discussion

The results of each model suggest the need for a high-quality neural translation approach on code-mixed data. Both of the LLMs used in the evaluation performed the best overall. Further analysis of the results of each LLM revealed that the Qwen model produces more standard and consistent native Sinhala output. It handles transliteration of romanized Sinhala to native Sinhala well by employing a more formal grammatical structure. On the downside, occasionally it has failed in translating English words to the appropriate Sinhala; instead, the model would perform transliteration. Since the dataset is created from social media data, the over-formalization has made the translation more unnatural compared to actual Sri Lankan digital discourse.

Compared to Qwen, the Gemini model exhibited superior transliteration of romanized Sinhala, yielding more natural translations that preserved the original semantic intent. While Gemini effectively identifies and transliterates named entities, it occasionally fails to provide full translations, instead generating English-Sinhala code-mixed outputs. However, evaluating Qwen and Gemini using few-shot prompting led to a significant performance increase compared to zero-shot baselines. The comparative evaluation results for both zero-shot and few-shot configurations are provided in the table 6.

The models fine-tuned on the SinMix2Mono training set performed better compared to others. Because the SinMix2Mono exposes models to real-world code-mixing patterns and informal linguistic designs, which are often absent in any other datasets. The fine-tuning process significantly improved the models' ability to distinguish between tokens requiring transliteration versus those requiring translation.

The lowest evaluation scores were given by the rule-based model, highlighting the limitation of

handling code-mixing in rule-based approaches. Apart from the Qwen, Gemini, and other fine-tuned models, other models were not capable of translating English words to Sinhala in the given code-mixed romanized Sinhala sentence. Instead of translating, it involved transliterating English words into Sinhala based on their phonetics. Yet it demonstrates relatively strong evaluation scores because Romanized Sinhala is the matrix language in this dataset, and existing NMT models are capable of handling it. Table 5 summarizes the results of each model evaluated on the golden dataset.

7.3 Transliteration Ambiguity dataset Evaluation

To further validate the models' performance, the newly constructed code-mixed transliteration ambiguity dataset was used to evaluate the best-performing models in each category. While the golden dataset evaluates general code-mixed romanized Sinhala translation and transliteration, the ambiguity dataset's primary focus is on evaluating the code-mixed transliteration ambiguity handling behavior of the given model. Specifically, the SwaBhasha trigram model, Gemini 2.5 flash-lite, and the Swa-Bhasha model fine-tuned on SinMix2Mono training data were used. The evaluation has been conducted on both the Sinhala transliteration ambiguity dataset introduced by Perera and Sumanathilaka (2025) and on the code-mixed transliteration ambiguity dataset introduced in this paper.

Model	BLEU	chrF	BERTScore
Sinhala Transliteration			
SwaBhasha trigram model	62.88	86.91	86.00
Gemini 2.5 Flash-Lite	56.55	81.26	84.00
XLM-Roberta FT	76.42	92.47	91.13
Swa-Bhasha model FT	77.64	92.85	92.37
Code-Mixed Transliteration			
SwaBhasha trigram model	19.92	55.03	61.00
Gemini 2.5 Flash-Lite	39.17	65.93	73.00
XLM-Roberta FT	44.93	70.63	74.98
Swa-Bhasha model FT	49.29	74.62	78.90

Table 2: Evaluation results on the ambiguity translations and transliterations datasets.

Results show that existing models perform poorly on code-mixed transliteration ambiguity compared to Sinhala transliteration ambiguity across all evaluation metrics.

8 Conclusion

This paper introduces the first gold-standard evaluation dataset on code-mixed transliteration ambiguity and general data, and the largest training dataset for code-mixed Romanized Sinhala-Sinhala translation and transliteration. The dataset is constructed from real-world social media data and curated through a comprehensive pipeline involving targeted data collection, preprocessing, and multi-stage annotation to ensure high quality and reliability. To validate the applicability of the proposed golden datasets, we conducted a systematic evaluation of nine systems. The results indicate that early-stage approaches, such as rule-based systems, exhibit limited capacity when handling complex and noisy code-mixed inputs. In contrast, NMT models and LLMs show greater robustness and adaptability, making them effective for non-standard and code-mixed language scenarios. Overall, this work highlights the importance of context-aware translation and transliteration for code-mixed Sinhala Romanized text, while providing a benchmark for future low-resource NLP research. The dataset can be accessed here: SinMix2Mono Dataset

Limitation

The parallel dataset consists only of code-mixed romanized Sinhala text and relevant native Sinhala translation and transliteration. The source text would be in Latin characters only, and the target text would be in Sinhala Unicode only. All emojis and words containing special characters have been removed during the pre-processing. All of the evaluations were conducted using a single GPU with 15 GB of memory. We were able to manually translate and transliterate approximately 25,000 sentences, with only 10% of those used for the golden dataset, which was evaluated and annotated by three annotators and proofread. Future work should focus on expanding the dataset on both scale and diversity.

Ethical Considerations

This work uses publicly available data and datasets to construct the proposed parallel corpus. To protect privacy, we removed usernames, phone numbers, and other personal information from sentences to anonymize user identifiers. The dataset is intended strictly for research and educational purposes.

References

Bhattacharjee, Soham, Baban Gain, and Asif Ekbal. 2024. Domain dynamics: Evaluating large language models in English-Hindi translation. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 341–354, Miami, Florida, USA, November. Association for Computational Linguistics.

Dharmasiri, Sachithya and TGDK Sumanathilaka. 2024. Swa bhasha 2.0: Addressing ambiguities in romanized sinhala to native sinhala transliteration using neural machine translation. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 241–246. IEEE.

Feinstein, Alvan R. and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Gwet, Kilem Li. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *The British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, may.

Jayakody, Ravindu and Gihan Dias. 2024. Performance of recent large language models for a low-resourced language. In *2024 International Conference on Asian Language Processing (IALP)*, pages 162–167.

Kugathasan, Archchana and Sagara Sumathipala. 2020. Standardizing sinhala code-mixed text using dictionary based approach. In *2020 International Conference on Image Processing and Robotics (ICIP)*, pages 1–6.

Kugathasan, Archchana and Sagara Sumathipala. 2021. Neural machine translation for Sinhala-English code-mixed text. In Mitkov, Ruslan and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 718–726, Held Online, September. INCOMA Ltd.

Kumbhar, Madhuri and Kalpana Thakre. 2024. Language identification and transliteration approaches for code-mixed text. *Journal of Engineering Science and Technology Review*, 17:63–70, 02.

Mudiyanselage, Anuja Dilrukshi Herath Herath and TG Deshan K Sumanathilaka. 2024. Tamzhi: Shorthand romanized tamil to tamil reverse transliteration using novel hybrid approach. *The International Journal on Advances in ICT for Emerging Regions*, 17(1).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Perera, Sandun Sameera and Deshan Sumanathilaka. 2025. Evaluating transliteration ambiguity in adhoc romanized sinhala: A dataset for transliteration disambiguation. In Mitkov, Ruslan and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2025)*, September.

Perera, Sandun Sameera, Lahiru Prabhath Jayakodi, Deshan Koshala Sumanathilaka, and Isuri Anuradha. 2025. Indonlp 2025 shared task: Romanized sinhala to sinhala reverse transliteration using bert. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 135–140.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Popović, M. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Ranasinghe, Tharindu, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2025. Sold: Sinhala offensive language dataset. *Language Resources and Evaluation*, 59(1):297–337.

Ransaka. 2023. Ransaka/sinhala-bert-medium-v2 · Hugging Face huggingface.co. https://huggingface.co/Ransaka/sinhala-bert-medium-v2.

Rathnayake, Himashi, Janani Sumanapala, Raveesha Rukshani, and Surangika Ranathunga. 2022. Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. *Knowl. Inf. Syst.*, 64(7):1937–1966, July.

Sheth, Rajvee, Himanshu Beniwal, and Mayank Singh. 2025. COMI-LINGUA: Expert annotated large-scale dataset for multitask NLP in Hindi-English code-mixing. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng,

editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7973–7992, Suzhou, China, November. Association for Computational Linguistics.

Shibli, G.M.S. et al. 2023. Automatic back transliteration of romanized bengali (banglish) to bengali. *Iran Journal of Computer Science*, 6(1):69–80.

Smith, Ian and Uthayasanker Thayasivam. 2019. Sinhala-english code-mixed data analysis: A review on data collection process. In *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, volume 250, pages 1–6.

Srivastava, Vivek and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In Xu, Wei, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors, *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online, November. Association for Computational Linguistics.

Srivastava, Vivek and Mayank Singh. 2021. HinGE: A dataset for generation and evaluation of code-mixed Hinglish text. In Gao, Yang, Steffen Eger, Wei Zhao, Piyawat Lertvittayakumjorn, and Marina Fomicheva, editors, *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 200–208, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Sumanathilaka, T.G.D.K., Ruvan Weerasinghe, and Y.H.P.P. Priyadarshana. 2023. Swa-bhasha: Romanized sinhala to sinhala reverse transliteration using a hybrid approach. In *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, pages 136–141.

Sumanathilaka, Deshan, Nicholas Micallef, and Ruvan Weerasinghe. 2024. Swa-bhasha dataset: Romanized sinhala to sinhala adhoc transliteration corpus. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 189–194.

Sumanathilaka, Deshan, Isuri Anuradha, Ruvan Weerasinghe, Nicholas Micallef, and Julian Hough. 2025a. Indonlp 2025: Shared task on real-time reverse transliteration for romanized indo-aryan languages. *arXiv preprint arXiv:2501.05816*.

Sumanathilaka, Deshan, Sameera Perera, Sachithya Dharmasiri, Maneesha Athukorala, Anuja Dilrukshi Herath, Rukshan Dias, Pasindu Gamage, Ruvan Weerasinghe, and YHPP Priyadarshana. 2025b. Swa-bhasha resource hub: Romanized sinhala to sinhala transliteration systems and data resources. *arXiv preprint arXiv:2507.09245*.

Team, Qwen. 2025. Qwen3: Think deeper, act faster, April.

UCSC. 2006. Real Time Unicode Converter. https://ucsc.cmb.ac.lk/ltrl/services/feconverter/t1.html.

Uthpala, D. K. and S. Thirukumaran. 2024. Sinhala-english code-mixed language dataset with sentiment annotation. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 184–188.

Wisal, Muhammad, Abbas Mustafa, and Umair Arshad. 2022. Cmrutu: Code mixed roman urdu (roman urdu and english) to urdu translator. In *2022 24th International Multitopic Conference (INMIC)*, pages 1–5.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

A Data and Annotation

This section contains information and statistics related to the tasks done in data collection and the initial data annotation process.

Algorithm 1 Code-mixed Romanized Sinhala Data Collection and Filtering Pipeline

Require: YouTube video ID v, English dictionary $\mathcal{E}$, Romanized Sinhala dictionary $\mathcal{S}$
Ensure: Filtered code-mixed comment set $\mathcal{C}$

```
Initialize empty list C ← ∅
Retrieve all comments Y from video v
for each comment c ∈ Y do
    if c contains only Latin characters then
        Tokenize c into words T
        Check presence of English words using E
        Check presence of Romanized Sinhala words using S
        if c contains at least one word from E and one word from S then
            Append c to C
        end if
    end if
end for
Return C
```

Category	Count
News & Law/Politics	5,405
Entertainment	1,153
Science & Technology	2,782
Business & Education	1,954
General	8,661
Total	**19,955**

Table 3: Distribution of collected data

Issue	Input	Expected	Generated
Ambiguity	wahinna wage yanne	වහින්න වගේ යන්නේ	වහින්න වැටුප යන්නේ
Selection	mama eyawa grade 10 idn danawa	මම එයාව 10 වසරේ ඉදන් දන්නවා	මම එයාව ග්‍රේඩ් 10 ඉදන් දන්නවා
Unsuitable Translation	meh pen eka mn godak kal use kara	මේ පෑන මන් ගොඩක් කල් පාවිච්චි කරා	මේ පෑන මන් ගොඩක් කල් පරිශීලනය කරා
Sentence Restructure	ada class yanne na	අද පන්ති යන්නේ නෑ	අද යන්නේ නෑ මම පන්ති

Figure 8: Examples of identified error categories on LLM-based initial annotation

B Dataset Statistics

This section further discusses and analyses statistics of the dataset proposed in this study.

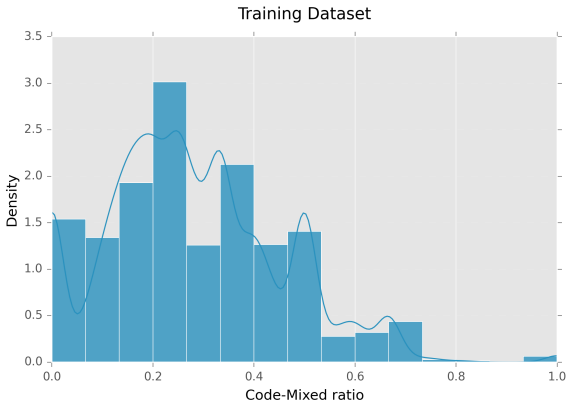

Figure 9: Distribution of English words ratios in Romanized Sinhala sentences in the Train dataset.

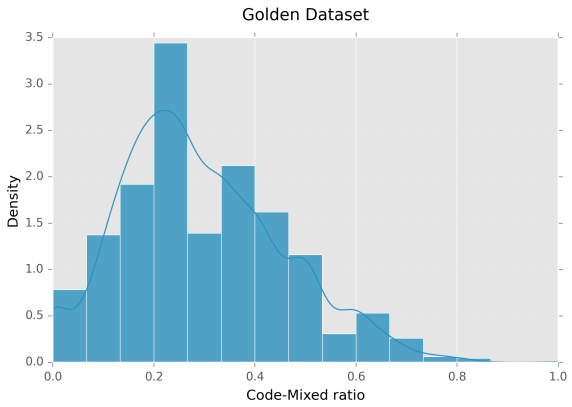

Figure 10: Distribution of English words ratios in Romanized Sinhala sentences in the Golden dataset.

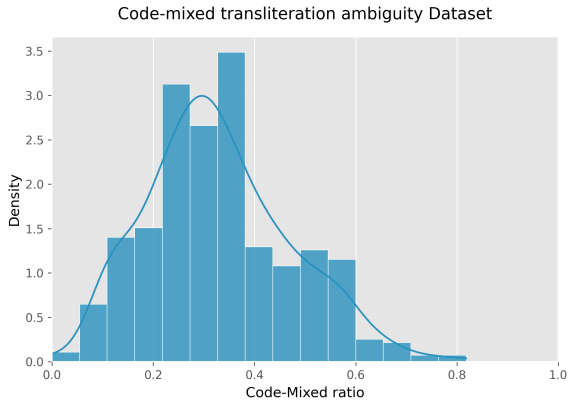

Figure 11: Distribution of English words ratios in Romanized Sinhala sentences in the code-mixed transliteration ambiguity dataset.

In both figure 9, 10, and 11, the left-skewed graph states that the dataset's matrix language is Romanized Sinhala and the embedded language is English. The peak density between 0.2 and 0.3 in the golden dataset indicates a code-mixing pattern of how English words are embedded in Romanized Sinhala.

POS Category	Golden	Train	Ambiguity
Noun (NOUN)	69.6%	66.9%	78.2%
Adjective (ADJ)	10.3%	10.2%	8.8%
Verb (VERB)	7.5%	7.4%	6.6%
Pronoun (PRON)	5.0%	6.2%	3.8%
Preposition (ADP)	2.2%	2.5%	1.4%
Adverb (ADV)	1.9%	2.1%	0.5%
Conjunction (CONJ)	0.9%	1.4%	0.3%
Numeral (NUM)	0.9%	1.1%	0.1%
Particle (PRT)	0.8%	0.9%	0.1%
Determiner (DET)	0.7%	1.0%	0.1%

Table 4: Part-of-Speech distribution of English words within Romanized Sinhala sentences

The Table 4 summarises the Part-of-Speech distribution on Romanized Sinhala sentences in Golden, Training and Ambiguity datasets. The Noun category dominates the highest ratio of English words. This confirms that this dataset follows the Matrix Language Frame.

Target Word	Romanized Sinhala	Native Sinhala	Context
pain	gedara langa nisa mama hama dama office ekata pain yanawa	ගෙදර ළඟ නිසා මම හැමදාම කාර්යාලයට පයින් යනවා	Sinhala
pain	gym giyapu nisa ada aga pathe okkoma pain	ජිම් ගියපු නිසා අද ඇඟ පතේ ඔක්කොම වේදනාව	English
pain	chest pain ekak thiyena kenekata pain yanna dena eka waradi	පපුවේ වේදනාවක් තියෙන කෙනෙකුට පයින් යන්න දෙන එක වැරදියි	Combined
pain	hitha riduna pain eka nisa mama painma gedara awa	හිත රිදුන වේදනාව නිසා මම පයින්ම ගෙදර ආවා	Combined

Figure 13: Example sentences in the Code-mixed Transliteration Ambiguity Dataset

The Figure 13 shows a few sentence samples for the word *'Pain'* for Sinhala, English, and combined contexts.

Prompt Used for code-mixed transliteration ambiguity dataset

A code-mixed romanized Sinhala to Sinhala translation transliteration system required to evaluate how the system handles WSD translation words.
To evaluate this a test dataset must be constructed. The selected WSD word is: '[word]'.

CONTEXT: - Sinhala Meaning: The word '[word]' is a transliteration of the Sinhala word '[sinhala word]'. - English Meaning: The word '[word]' also has these English meanings: [english context].

TASK: Generate exactly 15 unique sentences in Sri Lankan Singlish (Romanized Sinhala): 1. 5 sentences where '[word]' has ONLY the Sinhala meaning '[sinhala word]' (target sense: sin). 2. 5 sentences where '[word]' has ONLY the English meaning (target sense: eng). 3. 5 sentences where '[word]' appears TWICE using BOTH meanings (target sense: com).

STRICT RULES: - Use natural conversational Singlish. - Sentence length: 8–20 words. - The picked word must appear in the sentence. - Ensure the sentence is meaningful. - Sinhala outputs must contain only Sinhala Unicode.

[FEW-SHOT EXAMPLE]
Ensure that each sentence provides sufficient contextual information to help evaluate how well the translation transliteration system resolves ambiguity based on context.

Figure 14: Prompt used to generate the code-mixed transliteration ambiguity evaluation dataset

C Inter-Annotator Agreement

The following contains the Inter-Annotator Agreement that has been used in compiling the Golden dataset.

Annotation Process:
Romanized Sinhala section would/should have Latin or Romanized characters. The Sinhala section would/should have only Sinhala Unicode characters. The expected annotation is to perform translation and transliteration of Romanized Sinhala to Sinhala Unicode, while preserving the overall contextual meaning.

Annotation Guidelines:
1. English words should be translated into Sinhala.
2. Romanized Sinhala words should be transliterated to Sinhala.
3. The Sinhala words that are translated from English words should be the most used in daily communication. It is preffered to ignore more advanced and unused synonyms.
4. Named entities should be transliterated to

Ambiguous Word	English Meaning	Sinhala Word	Sinhala Meaning (Contextual)
path	An established line of travel or route	පත්	Leaves / Pages
wage	Regular payment for work; to carry on	වගේ	Like / Similar to
rate	Relative speed, rank, or cost	රටේ	In the country
game	A sport, contest, or animal hunted	ගමේ	In the village
bath	Act of washing the body; a vessel	බත්	Cooked rice
begin	To start or set in motion	බැගින්	Each / Per
math	Science of numbers and logic	මාත්	Me too
kale	A curly-leafed cabbage; money	කාලේ	Time / Era
gene	A segment of DNA; unit of heredity	ගැනේ	About / Regarding
mile	A unit of linear distance	මිලේ	Price
pile	A heap, collection, or stack	පිලේ	On the veranda / porch
pole	A long rod; an inhabitant of Poland	පොලේ	In the market
same	Identical; closely similar	සමේ	On the skin
dine	To eat dinner; to host for dinner	දිනේ	The date / day
hire	To employ, lease, or rent	හිරේ	In jail / prison
sale	Selling at reduced prices; general selling	සාලෙ	In the living room
siren	A warning sound; a seductive woman	සිරෙන්	From the jail / cell
pine	A coniferous tree; to desire	පිනේ	Of merit / Through luck
ride	To travel in a vehicle or on an animal	රිදේ	It hurts / It pains
ape	A large primate without a tail	අපේ	Our / Ours
gas	A fuel or substance in gaseous state	ගස්	Trees
name	A unit by which someone is known	නමේ	In the name
seen	Past participle of see; to perceive	සීන්	Scenes
madam	Polite way of speaking to a woman	මඩම්	Orphanages / Homes
papaya	A tropical fruit with orange flesh	පාපය	The sin
panama	A country on the isthmus	පණම	Life itself
mitten	A type of glove	මිටෙන්	From the handle / fist
pain	Physical or mental suffering	පයින්	On foot / Walking
rotten	Decaying or going bad	රොත්තෙන්	From the swarm / bunch
solos	Activities done alone	සොළොස්	Sixteen
message	Information sent to another	මැස්සාගේ	Of the fly / mosquito
bandage	Material used to bind a wound	බණ්ඩාගේ	Of Banda (personal name)
warden	A supervisor or guard	වරදෙන්	From the mistake / fault
hides	To put where it cannot be found	හිඩැස්	Gap / Space

Figure 12: List of code-mixed ambiguous words with relevant English meaning, Sinhala transliteration, and Sinhala meaning selected for code-mixed transliteration ambiguity dataset construction.

Sinhala, not translated.
5. If complex Roman sentences were identified (too many named entities, numbers, technical names(GPU, RAM, RTX)), remove the whole sentence, both singlish and Sinhalaa.

If something found that is not sure how to translate based on the given guidelines, highlight the cell and notify the owner of the dataset.

D Translation and Transliteration Prompt

The following contains the prompt used to obtain the translation and transliteration of the given source text from the Claude Sonnet 3.7 and 4 model.

This project contains a dataset for translating Sinhala Code-Mixed (SCM) text to proper Sinhala Unicode. The dataset consists of pairs of:
1.Romanized Sinhala text with Code-Mixed
2.Translated Sinhala Unicode text
- In the Sinhala column, it has provided a transliteration, but the English words should be translated to the equivalent Sinhala word. Consider the following examples [normal codemixed src and tgt example].
- Do not attempt to translate word by word. Consider the contextual meaning of the sentence. For example: [Word ambiguity examples and expected translations].
- When Named Entities are found, transliterate the text, do not translate. For example: [name entity example with expected translation].
- The translation of code-mixed romanized Sinhala should always be in Sinhala Unicode only. [List of code-mixed romanized Sinhala and its Sinhala translation examples].
- Based on the instructions, guidelines, and examples provided, update the translation/transliteration of Sinhala text relevant to the given source code-mixed romanized Sinhala.

Prompt for Few-shot evaluation.

You are an expert in Translation and Transliteration. Translate/transliterate the input Code-Mixed Romanized Sinhala sentences into natural monolingual Sinhala.

Rules:
1. Transliterate Romanized Sinhala words into Sinhala script.
2. Translate English words or phrases into natural Sinhala equivalents.
3. Do NOT keep English words in the output.
4. Preserve the SAME sentence order.
5. Return ONLY the Sinhala sentences. Do NOT include explanations.

[3 static few-shot examples]

Input: {Code-mixed Romanized Sinhala text}
Output:

Figure 15: Few-shot Prompt used for code-mixed Romanized Sinhala to Sinhala translation/transliteration.

E Golden Dataset Evaluation Results

Model	BLEU	chrF	BERTScore
Baseline Systems			
Rule-Based	4.03	37.40	40.00
SwaBhasha Trigram	17.35	51.65	57.00
Sinhala-BERT	22.28	54.31	59.00
Sinhala-BERT (finetuned)	21.64	54.05	59.00
Zero-shot LLMs			
Qwen-3 Max	18.47	55.87	60.03
Gemini 2.5 Flash-Lite	**26.36**	**61.91**	**66.00**
Few-shot LLMs			
Qwen-3 Max (few-shot)	20.57	**58.11**	62.00
Gemini 2.5 Flash-Lite (few-shot)	**33.43**	58.07	**71.00**
*Fine-tuned Models**			
XLM-R	42.33	71.52	76.58
M2M100	45.66	75.44	81.11
SwaBhasha-mBART	**49.70**	**78.47**	**81.19**

Table 5: Evaluation results on the Golden dataset using BLEU, chrF, and BERTScore metrics. *finetuned using Sin-Mix2Mono dataset

Model	Metric	Zero-shot	Few-shot	Δ
Qwen	BLEU	18.47	20.57	+2.1
	chrF++	55.87	58.11	+2.24
	BERTScore	60.03	62.00	+1.97
Gemini	BLEU	26.36	33.43	+7.07
	chrF++	61.91	58.07	-3.84
	BERTScore	66.00	71.00	+5.0

Table 6: Zero-shot vs few-shot performance for Gemini and Qwen across BLEU, chrF, and BERTScore. Δ shows the improvement from zero-shot to few-shot.