SiPaKosa: A Comprehensive Corpus of Canonical and Classical Buddhist Texts in Sinhala and Pali

Ranidu Gurursinghe♠ and Nevidu Jayatilleke♣
♠School of Computing, Informatics Institute of Technology, Sri Lanka
♣Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
ranidu.20222198@iit.ac.lk, nevidu.25@cse.mrt.ac.lk

Abstract

SiPaKosa is a comprehensive corpus of Sinhala and Pali doctrinal texts comprising approximately 786K sentences and 9.25M words, incorporating 16 copyright-cleared historical Buddhist documents alongside the complete web-scraped Tripiṭaka canonical texts. The corpus was created through high-quality OCR using `Google Document AI` on historical manuscripts, combined with systematic web scraping of canonical repositories, followed by rigorous quality control and metadata annotation. The corpus is organised into language-specific subcorpora: Sinhala and Mixed Sinhala-Pali. We evaluate the performance of language models using ten pretrained models, with perplexity scores ranging from 1.09 to 189.67 on our corpus. This analysis shows that proprietary models significantly outperform open-source alternatives by factors of three to six times. This corpus supports the pretraining of domain-adapted language models, facilitates historical language analysis, and aids in the development of information retrieval systems for Buddhist scholarship while preserving Sinhala cultural heritage.

Keywords: Sinhala NLP, Pali, Low-resource Language, Digital Humanities

1. Introduction

සබ්බදානං ධම්මදානං ජිනාති[1] (Thera and Rao, 1954), highlights that sharing knowledge and truth is considered the highest form of generosity because it provides the tools for others to free themselves. While the importance of the universal sharing of knowledge is highlighted, the field of computational linguistics remains significantly under-resourced in terms of structured religious corpora (Hutchinson, 2024).

The Sinhala language is part of the Indo-Aryan branch of the Indo-European language family with a rich and diverse literary heritage that has developed over several millennia. Its origins can be traced back to between the 3rd and 2nd centuries BCE. It is the primary language of the Sinhalese people, who represent the largest ethnic group in Sri Lanka, and it is recognised as the first language (L1) for approximately 16 million people (De Silva, 2026). Furthermore, Sinhala is classified as a lower-resourced language (Category 01) according to the criteria presented by Ranathunga and de Silva (2022).

Pali, which was historically a language commonly used by monks of different nations for communication (Zigmond, 2023), is considered a dead language despite being widely studied because it is the language of early Buddhist scriptures (Knauth and Alfter, 2014). In terms of data resources, Pali is also a category 1 low-resource language according to the criteria presented by Ranathunga and de Silva (2022). In this work, we will focus on Pali written in the Sinhala script (the Sinhala script is also used for writing Pali and Sanskrit literature in Sri Lanka (Gair, 1996)), rather than the more commonly known Devanagari script, which presents additional challenges.

The digitisation of historical texts presents several significant challenges. These include the difficulty of OCR on degraded copies, the inconsistency of spelling and orthography over the centuries (Jayatilleke and de Silva, 2025a), and the limited research that has been conducted on OCR for scanned documents and images featuring text written in the Sinhala script (De Silva, 2026).

We introduce SiPaKosa[2], a large-scale corpus combining classical Sinhala Buddhist texts from historical archives with complete canonical scriptures from web sources. Our contributions include: (1) a corpus of 786,839 sentences (∼9.25M words) from 16 historical documents and complete Tripiṭaka texts, (2) comprehensive metadata and language classification, (4) language-separated subcorpora Sinhala (465,539 sentences, 5.42M tokens), Pali (495 sentences, 3.2K tokens), and Mixed Sinhala-Pali (320,805 sentences, 3.83M tokens), and (5) evaluations showing 3-6 times gaps between proprietary and open-source models in terms of perplexity scores.

2. Related Work

2.1. The Bible in 100 languages

A multilingual Bible Corpus was introduced by Christodouloupoulos and Steedman (2015), comprising translations into 100 languages. The

[1] **English Meaning:** *The gift of dhamma (truth) conquers all other gifts; it truly excels and surpasses them all.*
Sinhala to IPA Transliteration: sʌbbʌðɑ:nʌŋ ðʰʌmmʌðɑ:nʌŋ dʒɪnɑ:θɪ.

[2] https://github.com/RanxduG/SiPaKosa-Dataset

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 68–84
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

corpus was constructed by web scraping publicly available Bible translations, automatic sentence alignment using the inherent verse-level structure of Biblical texts, and finally, quality filtering based on translation completeness and linguistic coverage. The resulting corpus structure preserves both the original hierarchical metadata (book names, chapter numbers, verse identifiers) and linguistic annotations (sentence boundaries, tokenisation), facilitating diverse research applications from machine translation to typological studies.

While serving different theological traditions, both the Bible and the Buddhist corpora share critical characteristics: these texts hold canonical status within their respective religious traditions (the Bible for Christianity, the Tripiṭaka for Buddhism). This status necessitates not only linguistic accuracy but also the preservation of religious meanings and theological concepts during digitisation. Furthermore, both face challenges related to translation and OCR across diverse linguistic families, along with the responsibility to safeguard sacred texts through digital means while ensuring accuracy for researchers who will utilise this data to study doctrinal texts.

2.2. Sinhala Diachronic Corpus

The *Sinhala Diachronic Corpus Version-1.0* (SiDiaC-v.1.0) was introduced by Jayatilleke and de Silva (2025a) and comprises 58,027 word tokens drawn from 46 literary works spanning the 5th to 20th century CE (426 CE to 1944 CE), sourced from the *National Library of Sri Lanka*. Text extraction was carried out using Google Document AI OCR, achieving an average accuracy of 96.84% across all processed documents. The corpus was carefully constructed to ensure balanced temporal coverage whilst prioritising canonical literary works; among the 46 books included, 18 are religious texts predominantly focused on Buddhism, annotated with a two-layer genre scheme (Fiction/Non-Fiction at primary level; Religious, History, Poetry, Language, and Medical at secondary level). SiDiaC-v.1.0 implements copyright filtering based on Sri Lanka's 70-year post-mortem *auctoris* rule under the Intellectual Property Act No. 36 of 2003, a practice directly adopted by SiPaKosa.

The second version of this corpus, SiDiaC-v.2.0 (Jayatilleke et al., 2026), substantially expands the resource to 241k words across 185 literary works, with publication dates ranging from 1800 CE to 1955 CE and written dates spanning the 5th to the 20th century CE. A date-annotated subset of 59 documents totalling approximately 67k words is further stratified by estimated written date, enabling fine-grained diachronic analysis. Of the 185 documents, 86 are classified under the religious genre, reflecting the enduring centrality of Buddhism to Sinhala literary culture. SiDiaC-v.2.0 addresses several limitations of its predecessor, including more comprehensive post-processing to correct malformed tokens, code-mixed content (Pali, Sanskrit, and English), and multi-column rendering errors introduced by OCR; it also broadens the written-date annotation methodology from strict manuscript-composition dating to the publication-year approach used in corpora such as COHA (Davies, 2012), enabling inclusion of a wider range of literary works.

SiDiaC focuses on diachronic language change across secular literary traditions and provides the most directly comparable Sinhala corpus to SiPaKosa in terms of construction methodology.

2.3. Religious and Cultural Text Corpora

The digitisation of religious texts in low-resource languages presents unique challenges that require not only linguistic accuracy but also cultural sensitivity and the preservation of specialised terminology. Several major religious corpora initiatives have established important precedents across traditions.

The Digital Pali Canon provides comprehensive coverage of Theravāda scriptures in Pali,[3] whilst the *BDK English Tripiṭaka*[4] offers scholarly English translations of the Chinese Tripiṭaka. The *Chinese Buddhist Electronic Text Association* (CBETA) represents one of the most comprehensive Buddhist digitisation efforts globally, providing over 100 million characters of Chinese Buddhist texts drawn from the Taishō Shinshū Daizōkyō (Volumes 1-55 and 85; 5,320+ individual texts) and supplementary collections (Wittern, 2002). Files are encoded in TEI∼P5 XML with Unicode handling for 2,800+ rare Buddhist characters proposed to the Unicode Consortium, released under CC BY-NC-SA 2.5∼TW.

The GRETIL (*Göttingen Register of Electronic Texts in Indian Languages*) project provides over 1,000 electronic editions primarily in Sanskrit, with additional texts in Pali, Prakrit, Tamil, Malayalam, Hindi, Marathi, Old Javanese, and Tibetan (Maas, 2020). The total collection spans several gigabytes in plain text, HTML, and TEI-conformant XML, released under CC∼BY via Zenodo. However, GRETIL provides minimal annotation (raw text only) and an ongoing migration from legacy encodings; whilst it includes some Pali texts, it provides no NLP-ready annotation layers, unlike SiPaKosa.

The *Quranic Arabic Corpus* (Dukes and Habash, 2010) provides morphological annotation of all 77,430 words of the Qur'an across 128,219 morphemic forms (clitic-level segmentation) using a

[3] https://tripitaka.online/

[4] https://www.bdkamerica.org/tripitaka-list/

Corpus	Language(s)	Tradition	Tokens	Units	Annotation	Alignment
Bible 100 (2015)	100 langs	Christian	varies	∼31K verses/lang	None	Verse
CBETA (2002)	Classical Chinese	Buddhist	100M+ chars	5,320+ texts	TEI XML	-
GRETIL (2020)	10+ Indic	Multi	GBs (raw)	1,000+ texts	Raw text	-
Quranic Arabic (2010)	Arabic	Islamic	77,430 words	6,236 āyahs	Morph+Syn+Sem	Verse
Tanzil (2007)	40+ langs	Islamic	varies	6,236 verses/lang	None	Verse
DCS (2020)	Sanskrit	Multi	2.5M+ items	∼650K sents	Lemma+POS+Morph	-
SansTib (2022)	Skt-Tibetan	Buddhist	14.4M tokens	317K pairs	Aligned	Sentence
MITRA (2026)	Skt+Zh+Tib+Pali	Buddhist	varies	1.74M pairs	MT-aligned	Sentence
SiDiaC-v.2.0 (2026)	Sinhala	Mixed	241K words	185 works	Genre+date	-
SiPaKosa [our work]	Sinhala + Pali	Buddhist	**9.25M tokens**	**786K sents**	Lang. class + meta	-

Table 1: Comparison of SiPaKosa with related religious and Indic text corpora.

44-tag POS tagset derived from traditional Arabic grammar, spanning 114 surahs and 6,236 āyahs. The corpus provides four annotation layers: morphological segmentation, part-of-speech tagging, syntactic dependency analysis, and semantic ontology, making it the most methodologically comparable sacred-text annotation project to what SiPaKosa could aspire to in future extensions. Building upon this foundation, the *Quranic Arabic Dependency Treebank* (QADT) (Dukes et al., 2010) extends morphological analysis with gold-standard syntactic annotation for approximately 37,578 words (∼49% of the Qur'an), achieving an F-measure of 78% with a rule-based dependency parser trained on the annotated data.

The *Tanzil parallel corpus* (Tanzil Project, 2007) offers Qur'anic translations in over 40 languages with verse-level alignment across 6,236 verses and 114 surahs, facilitating cross-lingual Islamic studies. The *Hebrew Bible Treebank* (Sade et al., 2018) offers morphological and syntactic annotation of Biblical Hebrew with detailed morphological features following traditional Hebrew grammar. The *Digital Corpus of Sanskrit* (DCS) (Hellwig et al., 2020) contains approximately 650,000 sentences with over 2.5 million lexical items annotated for lemmatisation, POS tagging, and morphological analysis, covering Sanskrit texts from 500 BCE to 1900 CE; later extensions added syntactic annotation in Universal Dependencies format (Hellwig and Nehrdich, 2018) and word sense annotation. At 650k sentences, the DCS is the most directly comparable single-language annotated ancient-text corpus to SiPaKosa's 786k sentences, and both address low-resource ancient languages; however, DCS does not cover Buddhist Pali in the Sinhala script, nor does it address language mixing between a living vernacular and a classical liturgical language.

These projects collectively demonstrate shared requirements: (1) domain expertise for doctrinal accuracy, (2) preservation of specialised religious terminology, (3) careful handling of language mixing between sacred languages (e.g. Pali-Sinhala, Arabic-Urdu, Sanskrit-Hindi), and (4) maintenance of cultural context. Table 1 presents a systematic comparison of SiPaKosa against related religious and Indic text corpora across six dimensions: language coverage, religious tradition, scale, annotation depth, and alignment granularity.

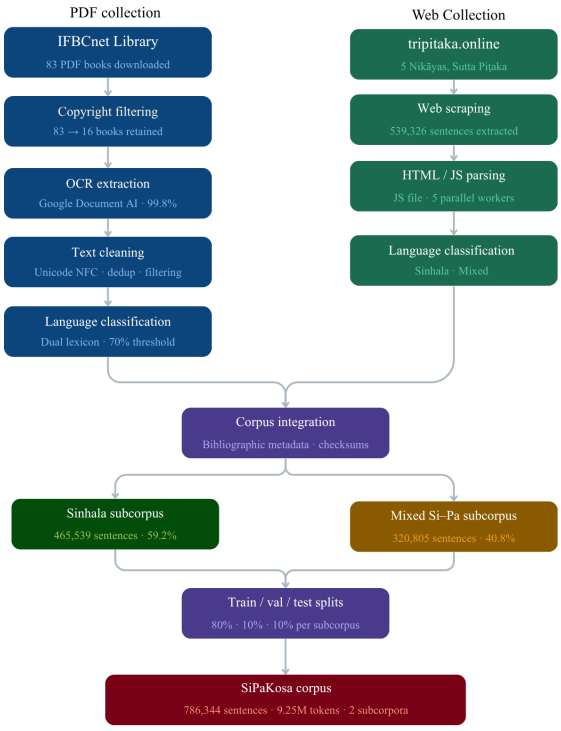

Figure 1: Complete methodology pipeline showing dual-source data collection, processing, and integration.

3. Methodology

We employ a dual-source methodology in order to cover canonical Buddhist scriptures with historically archived texts to achieve comprehensive coverage of Sinhala and Pali Buddhist literature. Our complete pipeline from data source identification through final corpus creation is illustrated in Figure 1. Furthermore, the book level filtering from a total of 83 to 16 eligible books is depicted in Figure 2.

3.1. Data Collection

Our historical corpus is compiled from the IFBCnet Library[5], a comprehensive digital archive of Sinhala and Pali Buddhist texts along with web-scrapping tripitaka.online[6], to complement historical texts with canonical scriptures. We downloaded 83 PDF books from IFBCnet Library, representing the complete available collection, each accompanied by bibliographic metadata including title, author information, publication year, and category classifications. The archive spans publications from the early days through the mid-20th century, providing valuable historical coverage of Sinhala and Pali Buddhist scholarship during a critical period of Buddhist revival and modernisation in Sri Lanka.

To ensure legal compliance, we implemented rigorous copyright filtering based on Sri Lanka's 70-year post-mortem auctoris rule under the Intellectual Property Act No. 36 of 2003[7] following the methodology by Jayatilleke et al. (2026). Firstly, Tripiṭaka canonical texts in PDF format were excluded via hardcoded rules, as these would be acquired separately through web scraping to ensure canonical completeness. Secondly, books authored by individuals who passed away before 1954 were classified as public domain with high confidence. Finally, books with living authors or incomplete author metadata were conservatively excluded to minimise legal risk. This conservative approach prioritises legal certainty over corpus size.

After copyright filtering, 16 books (19.3% of downloaded collection) were deemed suitable for inclusion. These books span three traditional Sinhala and Pali Buddhist literary categories: "Books related to Tipitaka" comprising 5 books (31.25%), which include canonical translations and commentaries; "Old Buddhist books" comprising 8 books (50.0%), which encompass philosophical treatises, devotional literature, and scholarly works; and "Buddhist characters" comprising 3 books (18.75%), which consist of hagiographies and biographical accounts of prominent Buddhist figures. The entire filtering process, from initial collection to the selection of books for the final corpus, is illustrated in Figure 2.

In addition to the historical PDF corpus, we systematically web-scraped tripitaka.online, a comprehensive digital repository of the Theravada Pali Canon maintained by Buddhist scholars, which provides Sinhala-language translations of the canonical scriptures alongside the original Pali

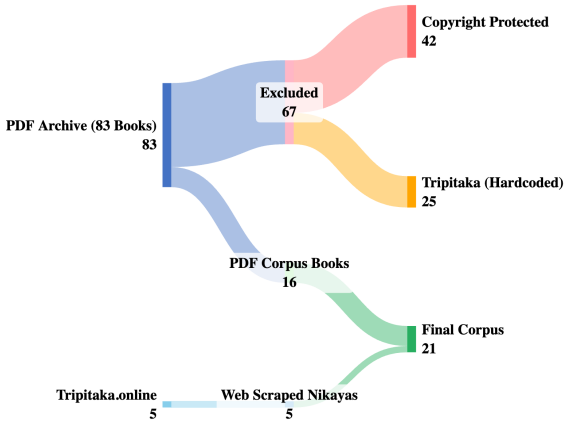

Figure 2: Book-level filtering from 83 downloaded books to 16 corpus-eligible books.

text. The Canon, known as the Tripiṭaka, literally *three baskets* in Pali, comprises the Vinaya Piṭaka (monastic discipline), the Sutta Piṭaka (discourses of the Buddha), and the Abhidhamma Piṭaka (systematic philosophical analysis). Our scraping focused on the five Nikayas of the Sutta Pitaka: Dīgha Nikāya (Long Discourses, 34 suttas), Majjhima Nikāya (Middle-Length Discourses, 152 suttas), Saṃyutta Nikāya (Connected Discourses, 65 suttas), Aṅguttara Nikāya (Numerical Discourses, 1,365 suttas), and Khuddaka Nikāya (Minor Collection, 831 suttas). Each sutta on tripitaka.online presents Pali canonical verses alongside corresponding Sinhala translations, providing natural parallel text alignment that preserves the semantic correspondence between source and target languages (this corpus lacks clear parallel text alignment but contains useful information for translation work). The detailed implementation is provided in Appendix B.1.

3.2. Text Extraction and Processing

We developed a two-stage extraction pipeline combining traditional PDF text extraction with OCR-based processing to handle the diverse quality of historical scans. For each PDF, we first attempted direct text extraction using pdfplumber, a Python library optimised for extracting text from PDF documents. As it did not produce precise extractions, we employed Google Document AI[8] (Jayatilleke and de Silva, 2025b), a production-grade OCR service with demonstrated strong performance on Indic scripts, including Sinhala (Jayatilleke and de Silva, 2025a).

This way, we ensure that if we can collect data from the PDF, we can get it without major issues or inconsistencies, but if the page is a scanned

[5] https://download.ifbcnet.org/

[6] https://tripitaka.online/

[7] https://www.gov.lk/wordpress/wp-content/uploads/2015/03/IntellectualPropertyActNo.36of2003Sectionsr.pdf

[8] https://cloud.google.com/document-ai/

image, we have to use the alternative route of OCR extraction.

For pages processed through `Document AI`, we implemented comprehensive quality control measures based on character-level confidence scores returned by the OCR engine. The OCR process achieved an average character-level confidence of 99.8% across all processed pages, demonstrating both the quality of the source scans and `Document AI`'s effectiveness on Sinhala script. Pages were classified into two confidence tiers: high confidence (mean confidence >=0.85, representing 85% of OCR-processed pages) and low confidence (mean confidence < 0.85, representing 15% of pages). Low-confidence pages were flagged for potential manual review and excluded from the primary corpus but preserved for future improvement efforts.

The proportion of pages achieving high confidence (85%) demonstrates both the quality of the archive scans and `Document AI`'s effectiveness on Sinhala script. We also implemented automated page classification to identify and filter non-content pages, such as covers, title pages, tables of contents, indices, and blank pages, using pattern-matching rules, retaining 7,064 content pages (99.8% retention).

For web-scraped canonical texts, we implemented a systematic crawler to extract structured `HTML` content from `tripitaka.online`, whilst preserving both Pali canonical text and Sinhala translations separately. After careful analysis, we identified that the data displayed on `tripitaka.online`, comes from a `js` (`JavaScript`) file, so we made sure the web scraper algorithm took this `js` file and read its content to get the textual data we needed, we preserved structural metadata during this process to organise the whole process which included nikaya identifiers, book numbers, sutta identifiers, and verse numbers for precise alignment and reference purposes. The extraction process required 72 hours, using multithreading with 5 parallel workers and rate-limiting constraints to ensure server stability, reflecting the substantial scale of the canonical corpus (539,326 sentences across five Nikayas).

3.3. Language Classification

Buddhist texts in Sri Lanka traditionally mix Sinhala and Pali with varying proportions depending on text type, historical period, and intended audience. To facilitate both monolingual and cross-lingual research, we implemented an automated language classification mechanism using lexicon-based matching. We constructed two primary lexicons: a Sinhala lexicon based on *Google's open-source Sinhala Pronunciation Lexicon*[9] containing 41,617 word forms covering a vast vocabulary, and a custom Pali lexicon containing 14,278 terms extracted from a Sinhala-Pali dictionary, which was included as one of the 16 books and then filtered for uniqueness to avoid overlap with Sinhala words. Further information on constructing the custom Pali Lexicon can be found at Appendix B.3

Pages and sentences were classified using a combined confidence scoring system that integrates lexicon-based word matching with morphological pattern analysis. For each text segment, we calculated a weighted confidence score for both Sinhala and Pali, where lexicon coverage (proportion of words matching each language's lexicon) received 70% weight and morphological feature detection (case markers, verbal endings, particles) received 30% weight. Text was classified as Sinhala only if the Sinhala confidence score reached at least 70% and exceeded the Pali confidence score. Text was classified as Pali if the Pali confidence score reached at least 70% and exceeded the Sinhala confidence score. Text was classified as Mixed if neither language's confidence score reached the 70% threshold, indicating substantial presence of both languages or ambiguous linguistic features. The 70% threshold was selected empirically after manual inspection of sample texts to balance precision (avoiding false classifications) and recall (capturing genuine monolingual content).

3.4. Corpus Integration and Metadata

We combined sentences from both sources into unified language-based subcorpora designed to support diverse research applications. The Sinhala subcorpus combines historical PDF Sinhala pages (111,632 sentences) with Tripitaka Sinhala translations (353,907 sentences), totalling 465,539 sentences representing purely Sinhala content suitable for monolingual language modelling and Sinhala-specific NLP tasks. The Mixed subcorpus combines historical PDF mixed pages (135,386 sentences) with Tripitaka Pali canonical text (185,419 sentences), totalling 320,805 sentences in both Sinhala and Pali that capture the language-mixing characteristic of Buddhist scholarly discourse.

For each book in the corpus, we collected comprehensive metadata in accordance with best practices in digital humanities corpus construction (Jayatilleke and de Silva, 2025a). The schema captures five categories of information: bibliographic details (source URLs, Sinhala and English titles, publication dates, and publisher information); author records (full names, dates of birth and death, and biographical notes); technical provenance (file

[9] https://raw.githubusercontent.com/google/language-resources/master/si/data/lexicon.tsv

sizes, SHA-256 checksums, download timestamps, and OCR confidence statistics); classification labels (category assignments, language classification results, and historical period); and copyright status (public domain eligibility and confidence level of assessment). The full metadata schema is provided in Figure 5.

We created train-validation-test splits at 80-10-10 ratios for each language subcorpus. For the Sinhala subcorpus, this resulted in 372,431 sentences for training, 46,553 sentences for validation, and 46,555 sentences for testing. For the Mixed subcorpus, this yielded 256,644 for training, 32,080 sentences for validation, and 32,081 sentences for testing. The training set comprises 629,075 sentences (80% of the total corpus), the validation set comprises 78,633 sentences (10%), and the test set comprises 78,636 sentences (10%) when aggregated across all language subcorpora and converted to sentence-level statistics. These splits are fixed and are released with the corpus to enable direct comparison of future research results.

4. Evaluation of SiPaKosa

4.1. Corpus Statistics

SiPaKosa comprises 786,839 total sentences from dual complementary sources: 247,513 sentences extracted from 16 copyright-cleared historical PDF books spanning 7,064 content pages, and 539,326 sentences from five Nikayas of web-scraped canonical texts. A comprehensive statistical analysis of the complete integrated corpus is presented in Table 2.

Source	Metric	Count
Historical PDF Corpus	Documents (books)	16
	Content pages	7,064
	Sentences	247,513
Web-Scraped Canonical Texts	Nikayas covered	5
	Sentences	539,326
Total	**Sentences**	**786,344**
	Tokens	**9,249,792**
	Average tokens/sentence	**11.76**

Table 2: Overall corpus statistics combining historical and canonical texts.

4.1.1. Language Distribution

The Sinhala subcorpus dominates at 59.2% of the total corpus (465,539 sentences), reflecting our focus on Sinhala Buddhist literature. The mixed subcorpus comprises 40.8% (320,805 sentences), capturing canonical Pali scriptures and mixed commentarial traditions. The statistics of each language subcorpus, revealing the complementary contributions of our dual sources, are detailed in Table 3.

The Sinhala subcorpus (465,539 total sentences) is predominantly web-scraped (76% from Tripitaka, 24% from PDF), reflecting the extensive online availability of canonical translations. The Mixed subcorpus (320,805 total) shows more balanced contributions (58% web-scraped, 42% PDF), capturing both canonical Pali verses and historical commentarial discourse. Based on our methodology for language classification in 3.3, there was a possibility of 3 language classes forming; unfortunately, no Pali-only content was found. The dual-source composition of each language subcorpus is illustrated in Figure 3.

Type	Metric	Sinhala	Mixed
Sentences	PDF corpus	111,632	135,386
	Web-scraped	353,907	185,419
	Total Count	**465,539**	**320,805**
	% of corpus	59.2%	40.8%
Tokens	**Total Count**	**5,418,716**	**3,831,076**
	% of corpus	58.6%	41.4%

Table 3: Sentence and token distribution by language subcorpus and source.

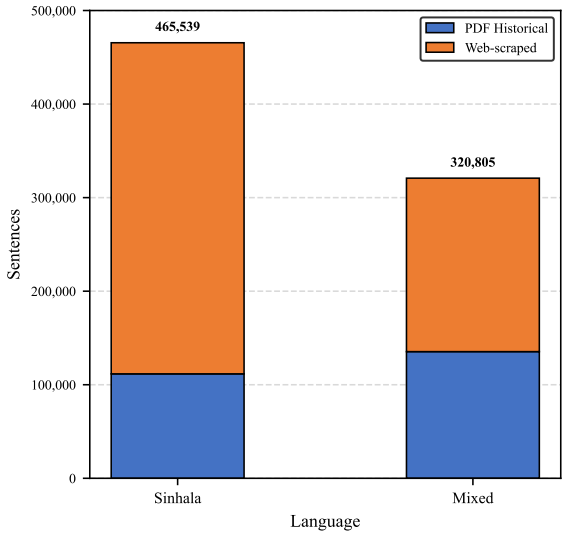

Figure 3: Sentence distribution by language and source.

At the token level, the corpus comprises 9,253,009 tokens, with the Sinhala subcorpus containing 5,418,716 tokens (58.6%) and the Mixed subcorpus containing 3,831,076 tokens (41.4%). The average sentence length varies across subcorpora: 11.6 tokens per sentence for Sinhala, 11.9 tokens per sentence for Mixed text.

4.2. Results

4.2.1. Evaluation Setup

We evaluate corpus quality and domain characteristics using nine state-of-the-art language models selected to represent diverse architectures, param-

Model	Buddhist Sinhala	Mixed Si-Pa	General Sinhala	BS/G Ratio	M/G Ratio
GPT-3.5-Turbo	**1.09**	**1.11**	**1.08**	1.01	1.03
GPT-4o-mini	1.23	1.15	1.31	0.94	0.88
GPT-4-Turbo	1.56	1.59	1.48	1.05	1.07
GPT-4o	1.62	1.77	1.68	0.96	1.06
Llama-3.1-8B-Instruct	3.29	4.18	3.26	1.01	1.28
Aya-Expanse-8B	6.21	8.82	4.90	1.27	1.80
Llama-3.2-3B-Instruct	6.62	7.81	6.94	0.95	1.13
Qwen2.5-3B-Instruct	7.24	9.81	6.56	1.10	1.49
Gemma-2-9B-It	22.21	36.71	14.85	1.50	2.47
SinLlama	183.33	189.67	101.57	1.80	1.87

Table 4: Comparison of perplexity among proprietary language models and small language models using different corpora.

eter scales, and training paradigms. Proprietary models include models from the GPT family; GPT-3.5-Turbo, GPT-4o-mini, GPT-4-Turbo, and GPT-4o (OpenAI, 2023). Open-source models include Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct (Grattafiori et al., 2024), aya-expanse-8b (Dang et al., 2024), Qwen2.5-3B-Instruct (Hui et al., 2024), and Gemma-2-9B-it (Team et al., 2024) with explicit multilingual and Sinhala support.

We evaluate models on three carefully constructed test sets designed to assess different aspects of Sinhala language modelling capability. The Buddhist Sinhala test set comprises 1,024 sentences sampled using diversity sampling from the pure Sinhala test split. The Mixed Sinhala-Pali test set comprises 1,024 sentences from the mixed test split, representing mixed Buddhist scholarly discourse. The General Sinhala test set comprises 1,024 sentences from a CulturaX[10], which consists of Sinhala news articles and online content. To ensure representative coverage while avoiding redundancy, we employ diversity sampling using K-means clustering (k=4) on multilingual sentence embeddings, selecting sentences closest to cluster centroids to capture diverse linguistic phenomena within each corpus.

We report corpus-level perplexity calculated as the exponentiated average negative log-likelihood across all tokens in each test set. For open-source models with direct access to model internals, given a test set with N tokens $w_1, w_2, \ldots, w_N$, perplexity is computed as;

$$\mathbf{PPL} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})\right) \quad (1)$$

where $P(w_i|w_{<i})$ is the model's predicted probability for token w_i given preceding context. For proprietary models that do not expose direct token-level log probabilities, we employ an approximate evaluation methodology in which models assess linguistic quality per sentence, with response confidence serving as a proxy for model certainty about the input text. Detailed proprietary model evaluation methodology is provided in Appendix C.1. To quantify domain-specific challenges independent of absolute perplexity values, we compute domain gap ratios as;

$$\mathbf{BS/G} = \frac{\mathrm{PPL}_{\mathrm{Buddhist-Sinhala}}}{\mathrm{PPL}_{\mathrm{General}}} \quad (2)$$

$$\mathbf{M/G} = \frac{\mathrm{PPL}_{\mathrm{Mixed}}}{\mathrm{PPL}_{\mathrm{General}}} \quad (3)$$

A comprehensive perplexity results across all nine models and three test corpora are presented in Table 4. The results reveal substantial performance variation between proprietary models and open-source alternatives.

4.2.2. Model Performance Analysis

GPT-3.5-Turbo achieves the best overall performance with the lowest perplexities of 1.09 for Buddhist Sinhala, 1.11 for Mixed Sinhala-Pali, and 1.08 for General Sinhala, demonstrating strong Sinhala language comprehension. All four proprietary models substantially outperform all locally-hosted open-source models, with GPT-3.5-Turbo, GPT-4o-mini, GPT-4-Turbo, and GPT-4o achieving perplexities between 1.08 and 1.77 across all test sets.

Among open-source models, Llama-3.2-3B-Instruct leads with perplexities of 3.29 for Buddhist and 4.18 for Mixed corpora, representing the best available option for resource-constrained settings. Aya-Expanse-8B, despite explicit multilingual training including Sinhala, achieves moderate performance (6.21 Buddhist, 8.82 Mixed), whilst

[10] https://huggingface.co/datasets/uonlp/CulturaX

smaller 3B models (`Llama-3.2-3B-Instruct`, `Qwen2.5-3B-Instruct`) demonstrate competent performance. `Gemma-2-9B-It` significantly underperforms with perplexities exceeding 14 on all corpora despite its 9B parameter count, indicating fundamentally inadequate Sinhala coverage in pretraining data.

Domain gap ratios reveal how models handle the specialised vocabulary and discourse patterns of Buddhist literature compared to general Sinhala. `GPT-3.5-Turbo` exhibits minimal domain gap with a BS/G ratio of 1.01 and M/G ratio of 1.03, suggesting a promising generalisation across registers and domains. `GPT-4o-mini` exhibits an inverse domain gap on Buddhist texts (BS/G ratio of 0.94), actually performing better on classical Buddhist Sinhala than general text. `Llama-3.1-8B` shows generalisation on pure Sinhala (BS/G ratio 1.01) but moderate difficulty with mixed text (M/G ratio 1.28). `Aya-Expanse-8B` and `Gemma-2-9B` exhibit larger domain gaps (1.27-2.47), with Gemma's extreme M/G ratio of 2.47 indicating the comprehension difficulty with language mixing.

Mixed Sinhala-Pali corpora present substantially elevated perplexity across most models, with an average M/G ratio of 1.42 across all models excluding the `Gemma` outlier. This consistent difficulty possibly reflects fundamental challenges in handling intra-sentential language mixing, in which Pali canonical quotations and Sinhala explanatory prose alternate rapidly. Proprietary models handle language mixing most effectively with M/G ratios ranging from 0.88 to 1.07. Open-source models struggle more significantly: `Llama-3.1` shows 28% perplexity increase (M/G ratio 1.28), `Aya-Expanse` shows 80% increase (1.80), and `Qwen2.5` shows 49% increase (1.49).

Another finding of this evaluation is the substantial performance gap between proprietary models and locally hosted open-source alternatives. `GPT-3.5-Turbo` outperforms the best local model `Llama-3.1-8B` by a factor of 3.0 on Buddhist Sinhala (perplexity 1.09 vs 3.29). This gap persists across all three test sets, suggesting it reflects fundamental differences in the scale and quality of the training data. The gap has critical practical implications: whilst proprietary models offer exceptional performance for production Buddhist NLP applications, they incur per-token costs and lack transparency. Open-source alternatives enable offline deployment, fine-tuning for specific Buddhist NLP tasks, and cost-effective experimentation.

Our findings for open source models challenge the common belief that larger models always outperform smaller ones, showing more complex relationships between parameter count and performance in low-resource languages. `Llama-3.2-3B` (perplexity 6.62) outperforms the larger `Aya-Expanse-8B` (6.21) despite having fewer than half the parameters, whilst `Gemma-2-9B`'s catastrophic performance (22.21) demonstrates that scale alone provides no guarantees. For proprietary models where accuracy is paramount, and API costs are acceptable, `GPT-3.5-Turbo` and `GPT-4o-mini` provide powerful performance. For research systems requiring local deployment and customisation, `Llama-3.1-8B` represents the best foundation despite its 3 times higher perplexity. In resource-constrained settings, `Llama-3.2-3B` or `Qwen2.5-3B` offer strong performance with manageable computational requirements.

5. Conclusion

We have presented SiPaKosa, a comprehensive corpus of canonical Sinhala and Pali Buddhist texts compiled from dual sources: historical public domain archives and systematically web-scraped canonical scriptures. Through rigorous digitisation, OCR processing, web scraping, copyright filtering, and quality assurance, we created a high-quality resource comprising 786,839 sentences (9.25 million tokens) across Sinhala, Pali, and mixed subcorpora, addressing critical gaps in Sinhala NLP while preserving important cultural and religious heritage.

Our comprehensive baseline evaluations across nine language models demonstrate significant performance variation, with proprietary models achieving remarkably low perplexities of 1.08 to 1.77 and substantially outperforming open-source alternatives with perplexities of 3.29 to 36.71. This performance gap highlights both the current state of Sinhala language modelling and opportunities for improvement in open-source multilingual systems. All models struggle with Sinhala-Pali mixed text, indicating opportunities for specialised model development targeting classical Buddhist literature.

This corpus serves as a critical resource for diachronic linguistic studies, providing insights into the history and evolution of language, as well as the profound influence of religious thought on everyday speech. In addition to formal canonical analysis, this corpus allows examination of folk speech and literary proverbs derived from classical Buddhist texts (Sofalas et al., 2026). This reflects the historical process by which scriptural concepts were integrated into the broader Sinhala linguistic identity.

Limitations

Temporal and Genre Coverage: Copyright restrictions exclude scholarly works published after the 1950s, potentially missing modern interpretations and contemporary Buddhist scholarship.

Post-Processing: Although the OCR extraction pipeline achieved an average character-level confidence of 99.8% across all processed pages, no manual post-processing was conducted on the extracted text. Correcting residual OCR errors and normalising historical spelling variations in classical Sinhala requires deep linguistic expertise in diachronic Sinhala and Pāli that was not available to the authors; consequently, such errors and orthographic inconsistencies may persist in the corpus.

Language Classification: Buddhist texts frequently mix Sinhala and Pali within sentences, making strict language separation imperfect. The Mixed category encompasses diverse text types, from balanced bilingual to predominantly Pali canonical texts. The 70% classification threshold is somewhat arbitrary and may benefit from optimisation. Additionally, this study did not explore potential Sanskrit text in these documents due to the absence of advanced language classification models.

Evaluation Limitations: Standard perplexity does not directly measure downstream task performance or semantic understanding. We lack human assessments of model outputs on Buddhist question answering or text generation tasks. Cannot analyse internal mechanisms of proprietary models or verify their Sinhala training data. Could evaluate other proprietary models rather than only OpenAI models

6. Bibliographical References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Salomey Osei, Shamsuddeen Hassan Muhammad, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

HWK Aravinda, Rashad Sirajudeen, Samith Karunathilake, Nisansa de Silva, Rishemjit Kaur, and Surangika Ranathunga. 2025. Sinllama-a large language model for sinhala. In *2025 Moratuwa Engineering Research Conference (MERCon)*, pages 617–622. IEEE.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Andreas Chari, Sean MacAvaney, and Iadh Ounis. 2025. Improving low-resource retrieval effectiveness using zero-shot linguistic similarity transfer. In *European Conference on Information Retrieval*, pages 290–306. Springer.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157.

Nisansa De Silva. 2026. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. BERTifying Sinhala - a comprehensive analysis of pre-trained language models for Sinhala text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7377–7385, Marseille, France. European Language Resources Association.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Kais Dukes, Eric Atwell, and Abdul-Baquee M. Sharaf. 2010. Syntactic annotation guidelines for the Quranic Arabic dependency treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Kais Dukes and Nizar Habash. 2010. Morphological annotation of Quranic Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier L de Lacalle, and Mikel Artetxe. 2024. Bertaqa: How much do language models know about local culture? *Advances in Neural Information Processing Systems*, 37:34077–34097.

James W Gair. 1996. Sinhala writing. *The world's writing systems*, pages 408–412.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2025. Cross-lingual transfer for low-resource natural language processing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.

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*.

Viktor Hangya and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2872–2885.

Oliver Hellwig and Sebastian Nehrdich. 2018. Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763, Brussels, Belgium. Association for Computational Linguistics.

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The treebank of vedic Sanskrit. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5137–5146, Marseille, France. European Language Resources Association.

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.

Zhiqi Huang, Hansi Hettiarachchi Ng, and Hamed Chen. 2023. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4561–4573.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Ben Hutchinson. 2024. Modeling the sacred: Considerations when using religious texts in natural language processing. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1029–1043, Mexico City, Mexico. Association for Computational Linguistics.

Nevidu Jayatilleke and Nisansa de Silva. 2025a. SiDiaC: Sinhala diachronic corpus. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pages 511–527, Hanoi, Vietnam. Association for Computational Linguistics.

Nevidu Jayatilleke and Nisansa de Silva. 2025b. Zero-shot OCR accuracy of low-resourced languages: A comparative analysis on Sinhala and Tamil. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 471–480, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Nevidu Jayatilleke, Nisansa de Silva, Uthpala Nimanthi, Gagani Kulathilaka, Azra Safrullah, and Johan Sofalas. 2026. SiDiaC-v.2.0: Sinhala diachronic corpus version 2.0. *arXiv preprint arXiv:2603.10861*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Jürgen Knauth and David Alfter. 2014. A dictionary data processing environment and its application in algorithmic processing of Pali dictionary data for future NLP tasks. In *Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing*, pages 65–73, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Dongyub Lee, Younghun Jeong, Hwa-Yeon Kim, Hongyeon Yu, Seunghyun Han, Taesun Whang, Seungwoo Cho, Chanhee Lee, Gunsu Lee, and Youngbum Kim. 2024. Tree-of-question: Structured retrieval framework for Korean question answering systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 406–418, Mexico City, Mexico. Association for Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Philipp André Maas. 2020. Pātañjalayogaśāstram. göttingen register of electronic texts in indian languages (gretil).

Sebastian Nehrdich. 2022. SansTib, a Sanskrit - Tibetan parallel corpus and bilingual sentence embedding model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6728–6734, Marseille, France. European Language Resources Association.

Sebastian Nehrdich and Kurt Keutzer. 2026. Mitra: A large-scale parallel corpus and multilingual pretrained language model for machine translation and semantic retrieval for p\= ali, sanskrit, buddhist chinese, and tibetan. *arXiv preprint arXiv:2601.06400*.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Soon Chang Poh, Sze Jue Yang, Jeraelyn Ming Li Tan, Lawrence Leroy Tze Yao Chieng, Jia Xuan Tan, Zhenyu Yu, Foong Chee Mun, and Chee Seng Chan. 2024. MalayMMLU: A multitask benchmark for the low-resource Malay language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 650–669, Miami, Florida, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+

questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. 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 823–848, Online only. Association for Computational Linguistics.

François Rémy, Julia Platzer, and Marco Dinarelli. 2024. Transtokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3456–3471.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sacred Texts Archive. 2020. The Internet Sacred Text Archive. Available at https://www.sacred-texts.com.

Shoval Sade, Amit Seker, and Reut Tsarfaty. 2018. The Hebrew Universal Dependency treebank: Past present and future. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Semantic Bible Project. 2015. The Semantic Bible: A multilingual semantic role annotation. Available at http://semanticbible.com.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Johan Nevin Sofalas, Dilushri Pavithra, Nevidu Jayatilleke, and Ruvan Weerasinghe. 2026. SinFoS: A parallel dataset for translating Sinhala figures of speech. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 8–26, Rabat, Marocco. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Tanzil Project. 2007. Tanzil quran text (version 1.1).

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Nārada Thera and B Veerabhadra Rao. 1954. *The Dhammapada*. John Murray London.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025a. MILU: A multi-task Indic language understanding benchmark. pages 10076–10132.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025b. MILU: A multi-task Indic language understanding benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10076–

10132, Albuquerque, New Mexico. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Christian Wittern. 2002. Chinese buddhist texts for the new millenium—the chinese buddhist electronic text association (cbeta) and its digital tripitaka.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

Wajdi Zaghouani, Abdelati Hawwari, and Mona Diab. 2012. A pilot PropBank annotation for Quranic Arabic. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 78–83, Montréal, Canada. Association for Computational Linguistics.

Dan Zigmond. 2023. Distinguishing commentary from canon: Experiments in pāli computational linguistics. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 213–222.

A. Additional Related Work

A.1. Multilingual Evaluation Benchmarks

Multilingual evaluation frameworks include XGLUE (Liang et al., 2020) (11 NLU and NLG tasks across 19 languages) and XTREME-R (Ruder et al., 2021) (10 tasks across 50 languages) for cross-lingual NLU tasks, building on English-only benchmarks GLUE (Wang et al., 2018) (9 NLU tasks, English only), SuperGLUE (Wang et al., 2019) (8 NLU tasks, English only), and SQuAD (Rajpurkar et al., 2016) (100,000+ extractive QA pairs, English only).

Language-specific MMLU variants have proliferated, covering Arabic (Koto et al., 2024), Chinese (Li et al., 2024), Turkish (Yüksel et al., 2024), Indonesian (Verma et al., 2025a), Korean (Son et al., 2025). Low-resource language benchmarks include AfriMMLU (Adelani et al., 2025) for African languages, MalayMMLU (Poh et al., 2024), BertaQA for Basque (Etxaniz et al., 2024), and MILU for 11 Indic languages (Verma et al., 2025b). The Include benchmark (Singh et al., 2025) attempts comprehensive multilingual coverage, though many benchmarks rely on automatic translation that can introduce errors and fail to capture cultural context (Adilazuarda et al., 2024).

Beyond the major religious corpus projects discussed in Sections 2.3-2.5, several specialised initiatives contribute to computational religious studies.

The Quranic PropBank (Zaghouani et al., 2012) extends semantic role labelling to Classical Arabic religious texts.

Christian Resources. The Semantic Bible project provides multilingual semantic role annotation for Biblical texts with ontological relationships (Semantic Bible Project, 2015).

Hindu and Vedic Resources. The Sansknet[11] project provides computational resources for Sanskrit including morphological analysers adapted to Vedic grammar.

Multi-Traditional Resources. The Sacred Texts Archive (Sacred Texts Archive, 2020) provides digitised versions of religious texts across 12 major traditions, though without linguistic annotation.

A.2. Indic Language Resources

SIPAKOSA exists within a vibrant Indic NLP ecosystem. Samanantar (Ramesh et al., 2022) is the largest publicly available Indic parallel corpus, comprising 49.7 million sentence pairs across 11 Indic languages; however, Pali is entirely absent from the collection, and the general-domain web-crawled text differs fundamentally from the specialised Buddhist canonical register that SIPAKOSA targets. IndicCorp v2 (Doddapaneni et al., 2023) provides 20.9 billion tokens across 24 Indic languages with the IndicXTREME benchmark (105 evaluation sets, 9 tasks); it includes Sinhala but not Pali, and its web-crawled sources contain no Buddhist domain content. The AI4Bharat-IndicNLP Corpus (Kunchukuttan et al., 2020) established 2.7 billion words for 10 Indic languages with pre-trained embeddings, providing foundational resources for the broader Indic NLP ecosystem.

These resources collectively provide broad coverage of general-domain Sinhala web text but offer zero coverage of Pali and zero domain-specific

[11] https://www.wilbourhall.org/sansknet/

Buddhist text, confirming that SiPaKosa addresses a genuine and unmet need within the Indic NLP landscape.

A.3. Language Identification

Since Sinhala and Pali share the same script (the Sinhala script is also used for writing Pali and Sanskrit in Sri Lanka; Gair 1996), language identification is a core technical challenge for SiPaKosa. Barman et al. (2014) established that approximately 7% of tokens in code-mixed text are ambiguous, a finding relevant to the Mixed Sinhala-Pali subcorpus.

A.4. Information Retrieval

Dense passage retrieval (Karpukhin et al., 2020) establishes neural retrieval baselines for open-domain question answering. For low-resource languages, optimal transport distillation (Huang et al., 2023) and zero-shot linguistic similarity transfer (Chari et al., 2025) improve cross-lingual IR effectiveness. Structured retrieval frameworks (Lee et al., 2024) provide QA methodologies that could be adapted for Sinhala Buddhist question answering.

A.5. Domain Adaptation

Parameter-efficient fine-tuning approaches, including LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023), enable domain adaptation without full fine-tuning, while knowledge distillation (Gou et al., 2021) supports creating smaller domain-specific models. These techniques are directly applicable to adapting general-purpose Sinhala models such as `SinBERT` (Dhananjaya et al., 2022) and `SinLlama` (Aravinda et al., 2025) for Buddhist domain tasks using SiPaKosa as the training corpus.

A.6. Cross-Lingual Transfer

García-Ferrero et al. (2025) provide comprehensive analysis of cross-lingual transfer techniques, whilst TransTokenization (Rémy et al., 2024) demonstrates vocabulary initialisation benefits for low-resource language adaptation of LLMs. Hangya and Fraser (2022) shows strategies for improving low-resource languages in multilingual models through vocabulary adaptation and continued pretraining, findings directly applicable to extending multilingual models for Sinhala Buddhist text.

A.7. LLM Evaluation

LLM evaluation surveys (Chang et al., 2024; Adlakha et al., 2024) discuss correctness and faithfulness assessment in instruction-following models, informing our baseline evaluation methodology. Hutchinson (2024) specifically addresses ethical and methodological considerations when using religious texts in NLP, including data provenance, cultural contexts, and proselytism concerns; this work directly informs SiPaKosa's commitment to copyright compliance and respectful representation of Buddhist canonical materials.

B. Additional Methodology

B.1. Web-scrapping

The web crawler for `tripitaka.online` implemented the following technical specifications: Rate limiting was set to 2 requests per second with exponential backoff retry logic (initial delay: 1s, maximum delay: 60s) to handle transient network failures. The crawler navigated the four-level hierarchy (Nikaya → Book → Sutta → Verse), extracted both Pali canonical text and Sinhala translations separately from the `JavaScript` file and preserved structural metadata, including nikaya identifiers, book numbers, sutta identifiers, and verse numbers for precise alignment. We employed a multi-threaded strategy with 5 parallel workers scraping 5 different suttas simultaneously, reducing total extraction time from an estimated 360 hours (sequential processing) to 72 hours (parallel processing), a 5× speedup. However, even with this optimisation, the extraction required 3 days of continuous operation due to the corpus scale (539,326 sentences) and necessary rate limiting (2 requests per second with exponential backoff) to avoid overwhelming the source server.

B.2. PDF Processing Flowchart

Figure 4 illustrates the end-to-end extraction pipeline applied to each of the 16 copyright-cleared PDF books. The pipeline begins by checking whether a book has already been processed, allowing interrupted runs to resume without reprocessing. Books that pass the 70-year post-mortem copyright check are then routed through one of two extraction paths depending on page content: digitally typed pages are handled by `pdfplumber`, while scanned pages are passed to `Google Document AI`, which returns character-level OCR confidence scores alongside the extracted text. All pages, regardless of extraction route, are classified into one of four structural categories: cover, table of contents, content, or index, and pages falling below the 0.7 confidence threshold are quarantined rather than discarded, preserving them for potential future review. Content pages that pass the confidence gate are written to disk as raw text, filtered to remove headers and OCR artefacts, and their provenance and quality metrics are recorded back into the per-book metadata record. The final output for

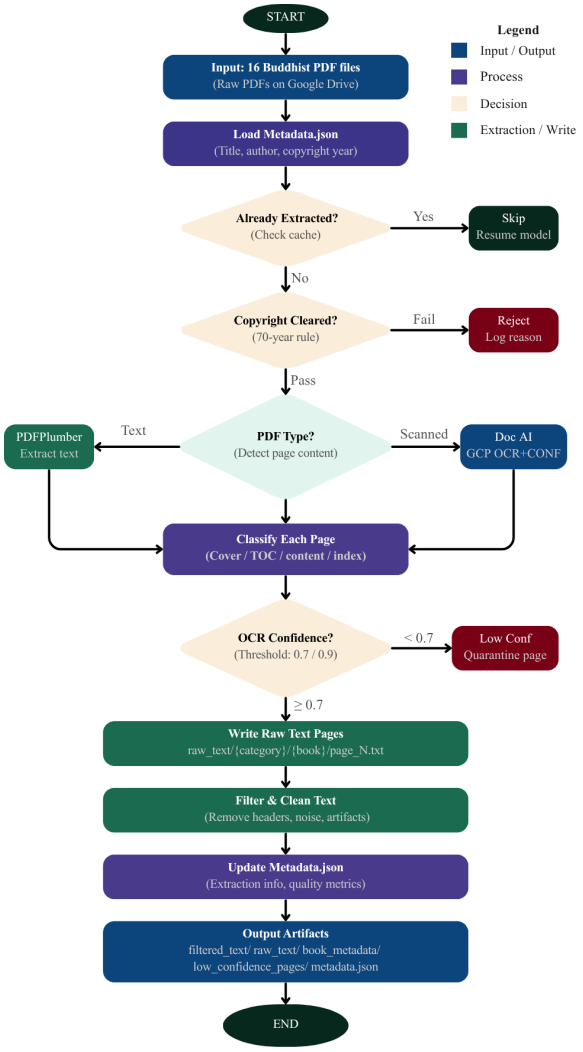

Figure 4: PDF processing pipeline applied to the 16 copyright-cleared historical Buddhist books, combining `pdfplumber` text extraction and Google Document AI OCR, with copyright checking, page classification, and confidence-based quality filtering (threshold: 0.7) prior to corpus integration.

each book comprises four artefact directories and an updated `metadata.json` file.

B.3. Custom Pali Lexicon

No publicly available Pali lexicon suitable for language classification existed at the time of this research, necessitating the construction of a custom resource. The source material was a Sinhala-Pali dictionary recovered as one of the 16 copyright-cleared books retained in the extraction pipeline described in Section 3.2 (Refer Table 5). All headword entries and their corresponding Pali equivalents were extracted from the dictionary's structured text output produced by the OCR stage, yielding a raw lexicon of 14,663 unique word forms. To isolate vocabulary that is diagnostically Pali rather than shared with modern Sinhala, the raw lexicon was filtered against a 41,617-word Sinhala lexicon obtained from Google's open-source language resources repository. Any token present in the Sinhala lexicon was removed from the Pali word list, on the basis that a word attested in the standard Sinhala vocabulary is not reliably diagnostic of Pali in a language classification context. The overlap analysis identified 887 shared words (6.05% of the raw Pali lexicon), predominantly short cognates of three to seven characters that are common to both languages due to their shared Indo-Aryan heritage. After removing these overlapping entries, the final filtered Pali lexicon comprised 13,776 unique word forms. This lexicon was then used alongside the Sinhala lexicon in the dual-lexicon classifier described in the following section to assign language labels at the paragraph and page level.

B.4. Corpus Metadata Information

```
{ "BaseURL": "str",
  "Category": "str",
  "Author": {
    "name": "str",
    "DOB": "str",
    "Date_of_passing": "str (YYYY-MM-DD)"
  },
  "book_name_si": "str",
  "book_name_en": "str",
  "number_of_pages": "int",
  "published_date": "str (YYYY-MM-DD)",
  "download_url": "str",
  "post_url": "str",
  "local_path": "str",
  "file_size_mb": "float",
  "sha256_checksum": "str",
  "download_timestamp": "str (ISO 8601)",
  "copyright_analysis": {
    "can_include_in_corpus": "bool",
    "reason": "str",
    "confidence": "str",
    "analysis_date": "str (ISO 8601)"
  }
{
  "extraction": {
    "status": "str",
    "extraction_timestamp": "str (ISO 8601)",
    "extraction_duration_seconds": "float",
    "page_info": {
      "total_pages": "int",
      "text_based_pages": "int",
      "ocr_pages": "int",
      "failed_pages": "int"
    },
    "methods_used": {
      "pdfplumber": "int",
      "document_ai": "int",
      "failed": "int"
    },
    "page_classification": {
      "cover": "int",  "toc": "int",
      "content": "int", "index": "int",
      "blank": "int",  "other": "int"
    },
    "quality_metrics": {
      "ocr_confidence": {
        "avg_confidence": "float",
        "min_confidence": "float",
        "max_confidence": "float",
        "low_confidence_pages": "list[int]",
        "distribution": {
          "high_0.9_to_1.0": "int",
          "medium_0.7_to_0.9": "int",
          "low_below_0.7": "int"
        }
      }
    },
    "file_paths": {
      "raw_text_dir": "str",
      "filtered_text_dir": "str",
      "book_metadata_dir": "str" }}}
```

Figure 5: JSON structure of a metadata record of the books in the IFBC corpus, defining the fields and their corresponding data types.

B.5. Copyright Filtered Books

The list of copyright-filtered books is listed in Table 5.

No.	Title	Author	Published
	Books Related to Tipitaka (5 books)		
1	Maithree Buddha Wanshaya nohoth anagatha wanshaya	Ven. Wilgammula Sangharaja Thero	2015
2	Milind Prashna	Ven. Heenatikubure Sumangala Nahimi	2015
3	Pretha Wasthu	Dhammapala Thero	2015
4	Vimana Wasthu	Dhammapala Thero	2015
5	Wishuddhi Margaya	Buddhaghosa Maha Thero	2015
	Old Buddhist Books (8 books)		
6	Jathila Thotilla	S. Mahinda Thero	2015
7	Pancha Maha Wadaya	Ven. Migettuwatte Gunananda Thero	2015
8	Pali Bhasha Shabdakoshaya	Ven. Widurupola Piyatissa Mahanayaka Thero	2015
9	Sinhala Deepavansaya	Unknown	2015
10	Buthsarana	Vidyachakrawarthi	2015
11	The Book of Dalada Pujavaliya	Unknown	2015
12	The Mahavamsa	Ven. Hikkaduwe Sri Sumangala Ther	2015
13	Saddharma Ratnavali	D.B. Jayatilaka	2015
	Buddhist Characters (3 books)		
14	Ven. Kadawadduwe Jinalankara Thero's Character	IFBC	2015
15	Ven. Man Buridaththa Thero's Character	IFBC	2015
16	Ven. Rerukane Chandawimala Maha Nayaka Thero's Character	IFBC	2015

Table 5: Copyright-cleared historical Buddhist texts included in the corpus (n=16). All books are in the public domain, with authors who passed away before 1954 (70+ years post-mortem).

C. Additional Result Information

C.1. Proprietary Model Evaluation

Proprietary models present a fundamental challenge for perplexity evaluation: production APIs typically do not expose token-level log probabilities required for exact perplexity calculation. Fortunately, OpenAI's API provides access to token-level logprobs through the `logprobs` parameter, enabling precise perplexity measurement for all four evaluated GPT variants.

C.1.1. OpenAI API Logprobs Implementation

For `GPT-3.5-Turbo`, `GPT-4o-mini`, `GPT-4-Turbo`, and `GPT-4o`, we utilise OpenAI's `logprobs` API parameter to extract token-level log probabilities for each input sentence. The evaluation process proceeds as follows:

1. For each test sentence, we make an API call with `logprobs=True` and `top_logprobs=1` to request token-level probability information.

2. The API returns log probabilities for each token in the input sequence (via the `prompt_logprobs` field in API responses where available, or through the `logprobs.content` structure in newer API versions).

3. We extract the log probability $\log P(w_i|w_{<i})$ for each token w_i in the sequence.

4. Perplexity is computed using the standard formula: $\text{PPL} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})\right)$.

5. We implement rate limiting (1-2 second delay between requests) to comply with API usage guidelines and avoid rate limit errors.

This methodology provides exact perplexity measurements for OpenAI models, identical in principle to the evaluation of open-source models with direct access to model internals. The primary difference is computational: API-based evaluation incurs per-token costs and requires network latency for each request, whilst local evaluation provides faster throughput at the cost of computational infrastructure requirements.

C.1.2. Model-Specific Considerations

`GPT-3.5-Turbo`: This provides the baseline for OpenAI's Sinhala capabilities. The API consistently provides logprobs access, enabling reliable perplexity measurement.
`GPT-4o-mini`: This lightweight variant of `GPT-4o` is optimised for cost-efficiency whilst maintaining strong multilingual capabilities. Logprobs access is consistent across API versions.
`GPT-4-Turbo`: The speed-optimised variant of `GPT-4` provides full logprobs access. Note that "Turbo" designation indicates architectural optimisations for inference speed rather than model capacity changes.
`GPT-4o`: The "omni" variant represents OpenAI's latest multimodal architecture with enhanced multilingual capabilities. Full logprobs access is provided, though API response structure differs slightly from earlier `GPT-4` variants (using `logprobs.content` rather than `prompt_logprobs`).

C.2. Proprietary vs Open-Source Models

The perplexity values obtained through OpenAI's API are directly comparable to those from open-source models evaluated locally, as both methods compute the identical mathematical quantity: the exponentiated average negative log-likelihood. The key advantages of API-based evaluation include:

- No local computational infrastructure required
- Access to proprietary models unavailable for local deployment
- Consistent evaluation environment (model version, ttokenisation inference configuration)

The primary disadvantages include:

- Per-token API costs accumulating across 1,024 sentences × 3 corpora
- Network latency and rate limiting, extending evaluation time
- Dependency on API availability and version stability
- Limited transparency regarding model architecture and training data